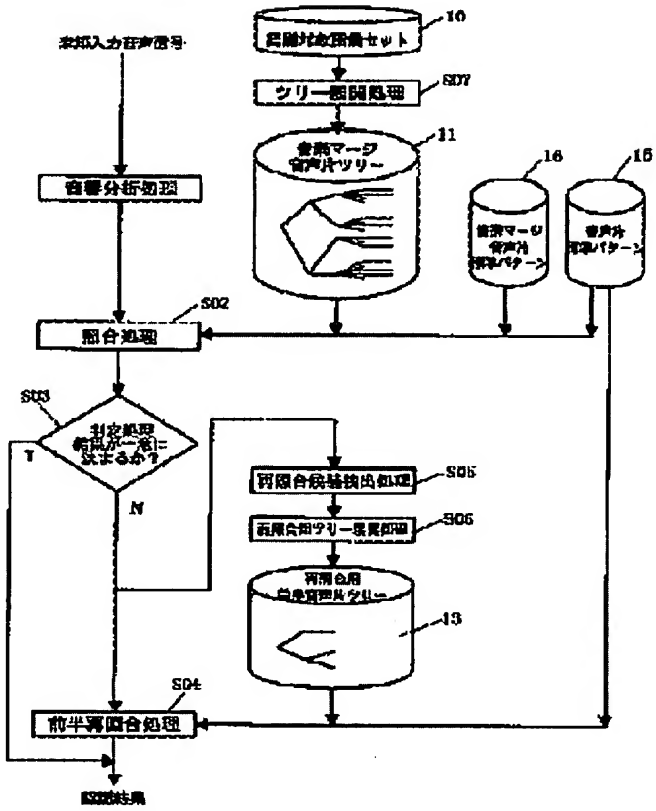
METHOD AND DEVICE FOR VOICE RECOGNITION, AND COMPUTER- READABL STORAGE MEDIUM

Publication number: JP2001312293
Publication date: 2001-11-09
Inventor: YAMADA MAKI; HOSHIMI MASAKATSU
Applicant: MATSUSHITA ELECTRIC IND CO LTD
Classification:
- International: ***G10L15/12; G10L15/08; G10L15/28;*** **G10L15/00;**
(IPC1-7): G10L15/12; G10L15/08; G10L15/28
- European:
Application number: JP20000129132 20000428
Priority number(s): JP20000129132 20000428

Abstract of **JP2001312293**

PROBLEM TO BE SOLVED: To recognize a voice with a small computation quantity without lowering the recognition performance as to a voice recognition technology. SOLUTION: This technology has a step wherein a phonetic notation series generated by merging phonemes having similar features of phonetic representation of an object vocabulary set to be recognized is converted into a voice segment series as a minimum unit of recognition and it is expanded into a phoneme merging voice segment tree, a collating step wherein previously found standard patterns representing features of a voice are connected according to the phoneme merging voice segment tree and collated with the feature vector time series of an unknown input voice signal by DP matching using a beam search over time matching, and a step wherein the standard patterns are connected according to a voice segment tree for re-collating if a result is not uniquely determined and then collated with the unknown input voice and the recognition result is outputted. Consequently, a voice can be recognized with a small computation quantity without lowering the recognition performance.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－312293
（P2001－312293A）
(43)公開日　平成13年11月9日(2001.11.9)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'（参考）
Ｇ１０Ｌ　15/12		Ｇ１０Ｌ　3/00	５３３Ｚ	５Ｄ０１５
15/08			５６１Ｊ	
15/28		5/06	Ｄ	

審査請求　未請求　請求項の数19　ＯＬ　（全 23 頁）

(21)出願番号　特願2000－129132(P2000－129132)

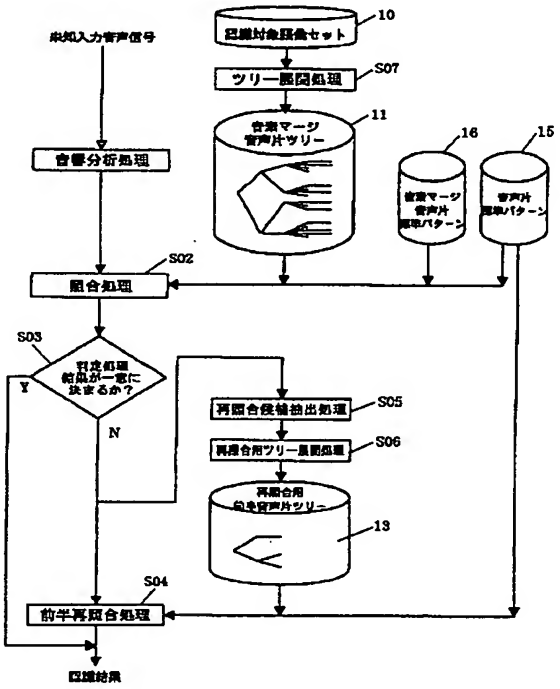
(22)出願日　平成12年４月28日(2000.4.28)

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地
(72)発明者　山田　麻紀
神奈川県川崎市多摩区東三田３丁目10番１号　松下技研株式会社内
(72)発明者　星見　昌克
神奈川県川崎市多摩区東三田３丁目10番１号　松下技研株式会社内
(74)代理人　100097445
弁理士　岩橋　文雄　（外２名）

Ｆターム(参考)　5D015 HH04 HH07 HH11 LL02

(54)【発明の名称】　音声認識方法およびその装置、並びにコンピュータ読み取り可能な記憶媒体

(57)【要約】
【課題】　本発明は音声認識技術に関するものであり、認識性能を落とすことなく少ない計算量で音声の認識を行うことを目的とする。
【解決手段】　認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換しこれを音素マージ音声片ツリーに展開するステップと、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行う照合ステップと、結果が一意に決まらなかった場合に、再照合用音声片ツリーに従って標準パターンを接続し、これと未知入力音声との照合を行い認識結果を出力するステップを有するもので、認識性能を落とすことなく少ない計算量で音声の認識を行うことができる。



1

【特許請求の範囲】

【請求項１】　未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するステップと、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行うステップと、照合結果が一意に決まる場合に認識結果を出力するステップと、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とする音声認識方法。

【請求項２】　音素マージ音声片ツリーに展開するステップは、認識対象語彙セットの音素表記の語頭から第Ｎ番目の音素までのうち特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開することを特徴とする請求項１記載の音声認識方法。

【請求項３】　音素のマージは、子音を音素群毎にまとめてマージすることを特徴とする請求項１または２記載の音声認識方法。

【請求項４】　未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第Ｎ番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第Ｎ番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行うステップと、照合の結果から再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とする音声認識方法。

【請求項５】　未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第Ｎ番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第Ｎ番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とする音声認識方法。

【請求項６】　精度を粗い音声片の標準パターンは、ラフ音声片１つの音声片あたりにかかる距離計算量を精密音声片１つあたりにかかる計算量に比べ削減することを特徴とする請求項４または５記載の音声認識方法。

【請求項７】　精度を粗い音声片の標準パターンは、認識結果が一意に決まる範囲内で、異なる音韻環境の音声片をマージすることを特徴とする請求項４または５記載の音声認識方法。

【請求項８】　音声片の距離計算量にかかるコストの削減は、ラフ音声片標準パターンのフレーム数を少なくすることを特徴とする請求項６記載の音声認識方法。

【請求項９】　音声片の距離計算にかかるコストの削減は、特徴パラメータベクトルの出現確率が複数のガウス分布の和、すなわちガウス分布の混合分布で近似できると仮定したとき、ラフ音声片標準パターンのガウス分布の混合数を少なくすることを特徴とする請求項６記載の音声認識方法。

【請求項１０】　音声片の距離計算にかかるコストの削減は、特徴パラメータベクトルの出現確率が複数のガウス分布の和、すなわちガウス分布の混合分布で近似できると仮定したとき、ラフ音声片標準パターンのガウス分布の共分散行列の種類数を少なくすることを特徴とする請求項９記載の音声認識方法。

【請求項１１】　再照合の際、未知入力音声の前半部分とのみ照合を行い認識結果を出力することを特徴とする請求項１、４、５のいずれかに記載の音声認識方法。

【請求項１２】　再照合の際、未知入力音声の発声区間すべてと照合を行い認識結果を出力することを特徴とする請求項１、４、５のいずれかに記載の音声認識方法。

【請求項１３】　未知入力音声の発声区間を特定せず、異なる始端を認める連続DPマッチングを用いたことを特徴とする請求項１、４、５のいずれかに記載の音声認識方法。

【請求項１４】　未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するツリー展開手段と、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パ

ターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行う照合手段と、照合結果が一意に決まるか否かを判定し、一意に決まる場合に認識結果を出力する判定手段と、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出する再照合候補抽出手段と、再照合用の音声片ツリーを展開する再照合用ツリー展開手段と、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力する再照合手段とを有することを特徴とする音声認識装置。

【請求項１５】　未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するツリー展開手段と、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行う照合手段と、照合の結果から再照合を行う候補となる認識対象語彙を抽出する再照合候補抽出手段と、再照合用の音声片ツリーを展開する再照合ツリー展開手段と、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力する再照合手段とを有することを特徴とする音声認識装置。

【請求項１６】　未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するツリー展開手段と、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力する照合手段とを有することを特徴とする音声認識装置。

【請求項１７】　プログラムされたコンピュータによって音声を認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記の

特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するステップと、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行うステップと、照合結果が一意に決まる場合に認識結果を出力するステップと、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体。

【請求項１８】　プログラムされたコンピュータによって音声を認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行うステップと、照合の結果から再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体。

【請求項１９】　プログラムされたコンピュータによって音声を認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビー

ムサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、ビームサーチを用いたDPマッチングを用いた音声認識方法およびその装置、並びにコンピュータ読み取り可能な記憶媒体に関するものである。

【０００２】

【従来の技術】認識対象となる音声の特徴を表現した標準パターンと、未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識する音声認識方法として、日本音響学会講演論文集,平成9年9月,3-1-4「音素の特徴点に着目した大語彙不特定話者音声認識法」が知られている。

【０００３】図１６は、従来の音声認識装置のコンピュータを用いた構成図を示すものである。図１６において、１は音声を取り込むマイク、２はＡ／Ｄ、３はインタフェース（Ｉ／Ｆ）、４はメモリ、５はＣＰＵ、６はキーボード／ディスプレイ、７はＣＰＵバス、８はＩ／Ｆ、９は出力、１０は認識対象語彙セット、１５は音声片標準パターン、１９は音声片ツリーである。

【０００４】上記のように構成された従来の音声認識装置の動作を図１７のフローチャートを用いて説明する。まず最初に、従来例における認識辞書にあたる音声片ツリー１９について、図３、図４を参照しながら説明をする。

【０００５】標準パターンの単位として、音素片、音素、音節、ＣＶ／ＶＣ（子音＋母音／母音＋子音）、ＶＣＶ、ＣＶＣなどが考えられる。これら認識の最小単位を音声片と呼ぶ。従来例では、子音の始端から母音中心までを表すＣＶと、母音中心から母音終端までを表すＶＣ、母音中心から母音中心までを表すＶＶを基本の単位とする。ＶＣは母音区間しか含まないが、後続子音により異なるＶＣと定義する。

【０００６】たとえば、認識対象語彙を「きりはら」「きりゅう」「ちり」「ちりゅう」「めぐろ」「めむろ」「ねむろ」「ふちゅう」の８単語としたとき、これらを音声片列で表すと、図４のようになる。

【０００７】これを単純なツリー構造で表したものが図３である。これを音声片ツリーと定義する。ここでは、アークに音声片を割り当てたが、ノードに割り当てることもできる。語彙の終端にあたるノードには、その語彙の終端であることがわかるようにしておく。このようなノードをリーフノードと定義する。図３ではリーフノードを黒丸で表している。また、ツリーの深さを、根から数えて第１段、第２段、…と数えるとする。

【０００８】以下、従来例について、図１７のフローチャートを参照しながらその動作を説明する。

【０００９】音声片標準パターン１５は、あらかじめ多数話者が発声した学習データから学習し、音声片毎に求めておく。本従来例では、特徴パラメータベクトルの出現確率が複数のガウス分布の和（これを混合分布と呼ぶ）で近似できると仮定し、学習データから、標準パターンのフレームごとにガウス分布の平均値ベクトルおよび共分散行列を求め、これを音声片標準パターン１５とする。

【００１０】音声片ツリー１９は、あらかじめ認識対象語彙セット１０から、ツリー展開処理Ｓ０６において作成しておく。

【００１１】まず、音響分析処理Ｓ０１は、入力された未知音声信号を分析時間（以下フレームと呼ぶ）毎にＮ個の特徴パラメータに変換される。特徴パラメータとしては、線形予測分析によるＬＰＣケプストラム係数、ＬＰＣメルケプストラム係数、メル線形予測分析によるメルＬＰＣケプストラム係数、メルスケールフィルタバンクによるメル周波数ケプストラム係数（ＭＦＣＣ）など、音声認識に適したものならばどのようなものを用いても良い。

【００１２】照合処理Ｓ０２では、音声片ツリー１９にしたがって音声片標準パターン１５を接続しながら、上記未知入力音声の特徴パラメータ時系列と標準パターンとの照合を行う。照合は、入力フレーム同期のビームサーチを用いたＤＰマッチングにより行う。照合の結果最も累積スコアの高かったリーフノードを求め、このリーフノードに対応する語彙を認識結果として出力する。

【００１３】以下に、ＤＰマッチングによる照合と、ビームサーチによる枝刈りのアルゴリズムについて説明する。

【００１４】ＤＰマッチングは、入力音声と標準パターンの時間整合をとりながら照合する方法である。第ｊ番目の入力フレームと、音声片ツリーの第ｋ番目のアークに対応する音声片ｍの標準パターンの第ｉ番目のフレームとの累積スコアＬ（ｉ，ｋ；ｊ）は、次の漸化式で表される。ただしｄ（ｉ，ｍ；ｊ）は入力の第ｊフレームと音声片ｍの標準パターンの第ｉフレームとの距離である。

【００１５】

【数１】

（初期化）
L (0.0;0) =0

(for j=1,…,J)

(for k=1,…,K)

L (0,k;j-1) =L ($I_{\hat{k}}$, $\hat{k}$;j-1)
$\hat{k}$は、アークｋの前に接続するアーク番号、$I_{\hat{k}}$はアーク$\hat{k}$の標準パターンの最終フレームを表す。

(for i=1,…,Ik)

$$L(i,k;j) = \max \begin{cases} L(i,k;j-1) + d(i,m;j) \\ L(i-1,k;j-1) + d(i,m;j) \end{cases}$$

【００１６】発声終了時に、リーフノードの累積スコア（終端アークの終端フレームにおける累積スコア）で最も大きいものが認識結果のスコアとなる。

【００１７】ビームサーチは、ＤＰマッチングの際スコアの低い経路は計算せずに、スコアの高い経路だけを伸張させながら計算していく手法である。累積スコアの低い経路であるかどうかは、入力と辞書の格子点における累積スコアの値が、その１フレーム前の時刻の最大累積スコアに比べ一定値（ビーム幅）以上低くなっているかどうかで判定する。累積スコアの低い格子点は枝刈りし、それ以外の格子点を候補として残していく。以下にそのアルゴリズムを示す。

【００１８】入力フレーム同期に、以下の式にしたがって、格子点（ｉ，ｋ；ｊ）を枝刈るか、候補として残すかの次の式によって判定をしながらＤＰパスを伸ばしていくものである。

【００１９】

【数２】

$$\theta = \max_{i,k} \{L(i,k;j-1)\} - \alpha$$

L (i,k;j) ≧θのとき、その格子点は残す。
L (i,k;j) ＜θのとき、その格子点は枝刈る。
ただし、αはビーム幅

【００２０】ビームサーチを用いたDPマッチングでは、発声開始付近ではまだどの候補も大きなスコアの差がつかないため、枝刈りはあまり行われない。そして発声後しばらくすると、発声内容とかけ離れた候補が枝刈りされはじめる。

【００２１】一方、認識対象語彙数が多い場合、語頭付近の音声片の種類数は非常に多くなる。そのため上記従来法の構成では、音声片ツリーは第１段目から大きく広がってしまい、発声開始付近では、照合のために非常に多くの経路について計算しなくてはならなくなってしまう。これはすなわち、発声開始付近では探索空間が広いと言える。

【００２２】したがって、発声開始付近では、探索空間が広い上に枝刈りがあまりなされないために、格子点候補数は爆発的にに増えてしまう。発声開始からしばらくすると、探索空間は広くても、枝刈りが多くなされるようになるため、格子点候補数は急激に減少する。

【００２３】従来法では、格子点候補数に比例して、認識にかかる計算量も増大する。したがって、従来法では認識にかかる計算量は図18のように時間変化する。図18を見てわかるとおり、発声開始付近での計算量は極端に多くなり、全体の計算量を削減するためには、発声開始付近での計算量を削減することが肝要である。

【００２４】単純に発声開始付近でのビーム幅を狭めることによっても発声開始付近での計算量を削減することはできるが、その場合正解候補が枝刈られやすくなる。発声開始付近では発声の言いよどみなどが起こりやすく、語頭のスコアが悪いというだけで枝刈りをしてしまうのは問題である。

【００２５】

【発明が解決しようとする課題】しかしながら上記の従来の構成では、発声開始付近で計算量が極端に多くなるという課題を有していた。

【００２６】本発明は、上記従来の課題を解決するもので、正解パスが枝刈られないようにしつつ語頭付近の探索空間を小さくするまたは語頭付近での照合にかかる計算量を削減する、すなわち認識性能を落とさずに全体の計算量を削減することを目的とする。

【００２７】

【課題を解決するための手段】この課題を解決するために、本発明は、音声片ツリーの語頭付近の広がりを狭めた音素マージ音声片ツリーを用いることによって格子点候補数を削減する、または音声片ツリーの語頭付近における音声片標準パターンの精度を粗くしたラフ音声片ツリーを用いることによって照合にかかる計算量を削減す

る。

【００２８】これは、ビームサーチで計算量の多い発声の前半部分は粗い照合を、ビームサーチで計算量の少なくなる後半部分は精密な照合をするという考え方に基づくものである。

【００２９】これにより、発声開始付近での計算量が削減し、認識性能を落とさずに計算量を削減することができる。

【００３０】

【発明の実施の形態】本発明の請求項1に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するステップと、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行うステップと、照合結果が一意に決まる場合に認識結果を出力するステップと、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有するものであり、音素をマージすることにより音声片ツリーの広がりが小さくなるため、探索空間が小さくなり1回目の照合にかかる計算量を大幅に削減することができ、再照合を行ったとしても全体の計算量を削減できるという作用を有する。さらに１回目の照合では音素マージを行うことにより特徴の似ている語彙は区別せずに認識を行うため1回目の照合で正解候補が漏れる可能性が低いという利点がある。

【００３１】請求項２に記載の発明は、請求項１記載の音声認識方法において、音素マージ音声片ツリーに展開するステップは、認識対象語彙セットの音素表記の語頭から第N番目の音素までのうち特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開することを特徴とするものであり、語頭から第N番目の音素までのうち特徴の似ている音素をマージすることにより、特に探索空間の広い語頭付近のツリーの広がりを狭めることができるため、探索空間が小さくなり1回目の照合にかかる計算量を大幅に削減することができ、再照合を行ったとしても全体の計算量を削減できるという作用を有する。さらに１回目の照合では音素マージを行うことにより特徴の似ている語彙は区別せずに認識を行うため1回目の照合で正解候補が漏れる可能性が低いという利点がある。

【００３２】請求項３に記載の発明は、請求項１または２記載の音声認識方法において、音素のマージは、子音を音素群毎にまとめてマージすることを特徴とするものであり、カテゴリー数が多く比較的認識が難しい子音を音響特徴の似通った音素群毎にまとめてマージするため、マージによる誤差が小さく認識性能を落とさずに効率よく計算量削減することができるという作用を有する。

【００３３】請求項４に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行うステップと、照合の結果から再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有するものであり、語頭付近での1回目の照合では音声片ツリーの前半部分は精度の粗いラフ音声片標準パターンを用いるため、1回目の照合にかかる計算量を大幅に削減することができ、再照合を行っても全体の計算量を削減できるという作用を有する。また再照合を行うことにより認識性能を落とさずに認識することができる。

【００３４】請求項５に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有するものであり、音声片ツリーの前半部分のみ精度の粗いラフ音声片標準パターンを用いて照合し、再照合をしないため、計算量は大幅に削減できるという作用を有する。音声片の標準パター

ンの精度を粗くするのは探索空間の広い語頭付近だけであるため、一律に音声片の標準パターンの精度を粗くするよりも効率的に計算量削減することができる。

【００３５】請求項６に記載の発明は、請求項４または５記載の音声認識方法において、精度を粗い音声片の標準パターンは、ラフ音声片１つの音声片あたりにかかる距離計算量を精密音声片１つあたりにかかる計算量に比べ削減することを特徴とするものであり、１つの音声片あたりにかかる距離計算量を削減する処理を設けることにより、容易に計算量を削減することができるという作用を有する。

【００３６】請求項７に記載の発明は、請求項４または５記載の音声認識方法において、精度を粗い音声片の標準パターンは、認識結果が一意に決まる範囲内で、異なる音韻環境の音声片をマージすることを特徴とするものであり、認識結果が一意に決まる範囲内で、異なる音韻環境の音声片をマージする処理を設けることにより、語頭付近の探索空間が狭まり効率よく計算量を削減することができるという作用を有する。

【００３７】請求項８に記載の発明は、請求項６記載の音声認識方法において、音声片の距離計算量にかかるコストの削減は、ラフ音声片標準パターンのフレーム数を少なくすることを特徴とするものであり、音声片標準パターンのフレーム数を削減する処理を設けることにより、容易に計算量を削減することができるという作用を有する。

【００３８】請求項９に記載の発明は、請求項６記載の音声認識方法において、音声片の距離計算にかかるコストの削減は、特徴パラメータベクトルの出現確率が複数のガウス分布の和、すなわちガウス分布の混合分布で近似できると仮定したとき、ラフ音声片標準パターンのガウス分布の混合数を少なくすることを特徴とするものであり、音声片標準パターンのガウス分布の混合数を削減する処理を設けることにより、容易に計算量を削減することができるという作用を有する。

【００３９】請求項１０に記載の発明は、請求項９記載の音声認識方法において、音声片の距離計算にかかるコストの削減は、特徴パラメータベクトルの出現確率が複数のガウス分布の和、すなわちガウス分布の混合分布で近似できると仮定したとき、ラフ音声片標準パターンのガウス分布の共分散行列の種類数を少なくすることを特徴とするものであり、音声片標準パターンのガウス分布の共分散行列を共通化する処理を設けることにより、容易に計算量を削減することができるという作用を有する。

【００４０】請求項１１に記載の発明は、請求項１、４、５のいずれかに記載の音声認識方法において、再照合の際、未知入力音声の前半部分とのみ照合を行い認識結果を出力することを特徴とするものであり、未知入力音声の前半部分とのみ照合を行い認識結果を出力する処理を設けることにより、再照合する区間が短くてすむため、再照合にかかる計算量を抑えることができるという作用を有する。

【００４１】請求項１２に記載の発明は、請求項１、４、５のいずれかに記載の音声認識方法において、再照合の際、未知入力音声の発声区間すべてと照合を行い認識結果を出力することを特徴とするものであり、未知入力音声の発声区間すべてと照合を行い認識結果を出力する処理を設けることにより、より精密な再照合が行えるため認識性能の劣化が少なくてすむという作用を有する。

【００４２】請求項１３に記載の発明は、請求項１、４、５のいずれかに記載の音声認識方法において、未知入力音声の発声区間を特定せず、異なる始端を認める連続DPマッチングを用いたことを特徴とするものであり、発声区間を特定しなくても、認識することができるという作用を有する。

【００４３】請求項１４に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するツリー展開手段と、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行う照合手段と、照合結果が一意に決まるか否かを判定し、一意に決まる場合に認識結果を出力する判定手段と、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出する再照合候補抽出手段と、再照合用の音声片ツリーを展開する再照合用ツリー展開手段と、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力する再照合手段とを有するもので、音素をマージすることにより音声片ツリーの広がりが小さくなるため、探索空間が小さくなり1回目の照合にかかる計算量を大幅に削減することができ、再照合を行ったとしても全体の計算量を削減できるという作用を有する。

【００４４】請求項１５に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するツリー展開手段と、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音

素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行う照合手段と、照合の結果から再照合を行う候補となる認識対象語彙を抽出する再照合候補抽出手段と、再照合用の音声片ツリーを展開する再照合ツリー展開手段と、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力する再照合手段とを有するもので、語頭付近での1回目の照合では音声片ツリーの前半部分は精度の粗いラフ音声片標準パターンを用いるため、1回目の照合にかかる計算量を大幅に削減することができ、再照合を行っても全体の計算量を削減できるという作用を有する。

【００４５】請求項１６に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するツリー展開手段と、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力する照合手段とを有するもので、音声片ツリーの前半部分のみ精度の粗いラフ音声片標準パターンを用いて照合し、再照合をしないため、計算量は大幅に削減できるという作用を有する。

【００４６】請求項１７に記載の発明は、プログラムされたコンピュータによって音声を認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するステップと、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行うステップと、照合結果が一意に決まる場合に認識結果を出力するステップと、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体であり、コンピュータに読み込み実行するものであり、音素をマージすることにより音声片ツリーの広がりが小さくなるため、探索空間が小さくなり1回目の照合にかかる計算量を大幅に削減することができ、再照合を行ったとしても全体の計算量を削減できるという作用を有する。

【００４７】請求項１８に記載の発明は、プログラムされたコンピュータによって音声を認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行うステップと、照合の結果から再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体であり、コンピュータに読み込み実行するもので、語頭付近での1回目の照合では音声片ツリーの前半部分は精度の粗いラフ音声片標準パターンを用いるため、1回目の照合にかかる計算量を大幅に削減することができ、再照合を行っても全体の計算量を削減できるという作用を有する。

【００４８】請求項１９に記載の発明は、プログラムされたコンピュータによって音声を認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたＤＰマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体であり、コンピュータに読み込み実

行するもので、音声片ツリーの前半部分のみ精度の粗いラフ音声片標準パターンを用いて照合し、再照合をしないため、計算量は大幅に削減できるという作用を有する。

【００４９】以下、本発明の実施の形態について図を用いて説明する。

【００５０】（実施の形態１）図１は、本発明の実施の形態１における音声認識装置のブロック構成図であり、以下に説明する。

【００５１】図１において、１は音声を取り込むマイク、２はＡ／Ｄ、３はインタフェース（Ｉ／Ｆ）、４はメモリ、５はＣＰＵ、６はキーボード／ディスプレイ、７はＣＰＵバス、８はＩ／Ｆ、９は出力、１０は認識対象語彙セット、１１は音素マージ音声片ツリー、１２はラフ音声片ツリー、１３は再照合用前半音声片ツリー、１４は再照合用音声片ツリー、１５は音声片標準パターン、１６は音素マージ音声片標準パターン、１７はラフ音声片標準パターン、１８は精密音声片標準パターンである。

【００５２】まず最初に、実施の形態１における認識辞書に当たる音素マージ音声片ツリー１１について、図３、図４、図５を参照しながら説明をする。

【００５３】標準パターンの単位として、音素片、音素、音節、ＣＶ／ＶＣ（子音＋母音／母音＋子音）、ＶＣＶ、ＣＶＣなどが考えられる。これら認識の最小単位を音声片と呼ぶ。本実施の形態では、子音の始端から母音中心までを表すＣＶと、母音中心から母音終端までを表すＶＣ、母音中心から母音中心までを表すＶＶを基本の単位とする。ＶＣは母音区間しか含まないが、後続子音により異なるＶＣと定義する。

【００５４】たとえば、認識対象語彙を「きりはら」「きりゅう」「ちり」「ちりゅう」「めぐろ」「めむろ」「ねむろ」「ふちゅう」の８単語としたとき、これらを音声片列で表すと、図４のようになる。

【００５５】これを単純なツリー構造で表したものが図３である。本実施の形態では、これを基本音声片ツリーと定義する。これは従来例で用いている音声片ツリーと同じものである。ここでは、アークに音声片を割り当てたが、ノードに割り当てることもできる。語彙の終端にあたるノードには、その語彙の終端であることがわかるようにしておく。このようなノードをリーフノードと定義する。図３ではリーフノードを黒丸で表している。また、ツリーの深さを、根から数えて第１段、第２段、…と数えるとする。

【００５６】音素マージ音声片ツリー１１は、基本音声片ツリーのうち第１段～第ｎ段までの音素をマージすることにより、語頭のツリーの広がりを小さくしたものである。第ｎ＋１段以降は基本音声片ツリーそのままである。

【００５７】第１段～第ｎ段までの音素マージは以下の方法で行う。日本語の母音は、５種類しかなくこれらを識別することは比較的容易であるが、子音はカテゴリー数も多く識別が難しい。そこで、子音は音素群（無声破裂音、摩擦音、有性破裂音など）毎にまとめてマージし、同じ音素群内の子音は区別をしないとする。すなわち、子音は、音素の区別は行わず、無声破裂音や摩擦音のような音素群の区別しか行わない。語頭音素が１文字違うだけの「きりゅう」と「ちりゅう」は区別を付けずに照合することになる。

【００５８】音響特徴の似通った子音の音素群内でのマージを行うため、マージによる誤差が少なく、しかも異なる音素群間の識別は音響特徴がかけ離れているため容易である。したがって、正解候補が枝刈られることはほとんどなく認識性能の低下が少ない。本実施の形態では、子音を図５のような４つのカテゴリーに分ける。

【００５９】音素をマージするとことにより音声片もマージされる。ＣＶは後続母音が同じ場合に、ＶＣは先行音素が同じ場合にマージする。音素群毎に音素をマージして得られる音声片を音素マージ音声片と定義する。音素マージ音声片のマージ方法と表記法の例を図６に示す。

【００６０】基本音声片ツリーのうち、第１段～第ｎ段までの音声片を、音素マージ音声片とすることにより、同じ音素マージ音声片を割り当てられたアーク同士をマージして語頭付近の広がりの小さいツリーにすることができる。これが音素マージ音声片ツリーである。

【００６１】図３の基本音声片ツリーを、第１段～第３段（ｎ＝３）までの音声片をマージして音素マージ音声片ツリーに変換したのが図７である。図７の音素マージ音声片ツリーは、図３の音声片ツリーに比べ、語頭付近のツリーの広がりが狭くなっている。ｎを１とすると語頭の１番目の音声片だけをマージしたことになり、∞とするとすべての音声片をマージすることになる。ｎの大きさは、計算量がリアルタイムで収まる程度に決めておくと効率が良い。音素マージ音声片ツリーでは、「きりゅう」と「ちりゅう」のようにリーフノードに複数の語彙が割り当てられることがある。

【００６２】次に、本発明の実施の形態１における音声認識装置について、図２のフローチャートを参照しながらその動作を説明する。

【００６３】図２において、音声片標準パターン１５は、あらかじめ多数話者が発声した学習データから学習し、音声片毎に求めておく。また、音素マージ音声片標準パターン１６は、マージする音声片すべての学習データから学習することにより求められる。たとえば、音声片／｛ｐ，ｔ，ｋ，ｃ｝ｉ／に対する標準パターンは、／ｐｉ／（ピ）、／ｔｉ／（ティ）、／ｋｉ／（キ）、／ｃｉ／（チ）のすべての学習データから学習することにより得られる。これを、あらかじめすべての音素マージ音声片について求めておくものとする。

【００６４】本実施の形態では、特徴パラメータベクトルの出現確率が複数のガウス分布の和（これを混合分布と呼ぶ）で近似できると仮定し、学習データから、標準パターンのフレームごとにガウス分布の平均値ベクトルおよび共分散行列を求め、これを標準パターンとする。

【００６５】音素マージ音声片ツリー１１は、あらかじめ認識対象語彙セット１０から、ツリー展開処理Ｓ０７において作成しておく。

【００６６】まず、音響分析処理Ｓ０１は、入力された未知音声信号を分析時間（以下フレームと呼ぶ）毎にＤ個の特徴パラメータに変換される。特徴パラメータとしては、線形予測分析によるＬＰＣケプストラム係数、ＬＰＣメルケプストラム係数、メル線形予測分析によるメルＬＰＣケプストラム係数、メルスケールフィルタバンクによるメル周波数ケプストラム係数（ＭＦＣＣ）など、音声認識に適したものならばどのようなものを用いても良い。

【００６７】照合処理Ｓ０２では、音素マージ音声片ツリー１１にしたがって音素マージ音声片標準パターン１６および音声片標準パターン１５を接続しながら、音響分析処理Ｓ０１からの未知入力音声の特徴パラメータ時系列と標準パターンとの照合を行う。照合は、入力フレーム同期のビームサーチを用いたＤＰマッチングにより行う。ＤＰマッチングの方法およびビームサーチの方法は、従来例と同じであるため説明を省略する。この照合を一回目の照合と呼ぶ。

【００６８】なお、本実施の形態では、音素マージ音声片ツリー１１はあらかじめ作成しておくとしたが、ビームサーチＤＰを行いながら動的にツリー展開してもよい。

【００６９】入力フレーム同期のビームサーチを用いたＤＰについて、その概念図を表したものが図９である。図９において、横軸は入力音声のフレーム、縦軸は音素マージ音声片ツリーにしたがって接続した音声片標準パターンのフレームを表している。辞書である縦軸はツリー状になっている。入力音声とツリー状の辞書のＤＰマッチングは、図９のようなツリー状のＤＰ面上での入力と標準パターンの最適な経路を求めながらスコアを算出するものである。このツリー状のＤＰ面は、第１段～第ｎ段までが音素マージされており枝の広がりは小さくなっている。

【００７０】ＤＰマッチングは、ビームサーチにより入力フレーム同期にＤＰ経路の枝刈りを行う。ビーム内に残る格子点候補数はＤＰ面のすべての格子点数に比べはるかに少ないため、このＤＰ面は実際にメモリ上に持つ必要はなく、仮想的なものである。

【００７１】発声開始からしばらくすると、発声内容と似ていない辞書のＤＰパスの累積スコアは、正解パスの累積スコアに比べ十分小さな値になり枝刈られるため、格子点候補数は急激に減少する。したがって、それまでの間の格子点候補数を抑えることが全体の計算量削減につながる。第一の実施の形態のように語頭付近のツリーの広がりを抑えることにより、発声開始付近のビーム内に残る格子点候補数は大幅に削減することができる。

【００７２】判定処理Ｓ０３では、ＤＰマッチングにより最も累積スコアの高かったリーフノード（最大ゆう度リーフノード）を求め、これに対応する語彙が一意に決まるかどうかの判定を行う。もし、一意に決まる場合（Ｙ）、すなわち最大ゆう度リーフノードに対応する語彙が１個しかない場合は、その語彙を認識結果として出力する。もし、一意に決まらない場合（Ｎ）、すなわち最大ゆう度リーフノードに対応する語彙が複数存在する場合には、次のような方法で認識結果を決定する。

【００７３】再照合候補抽出処理Ｓ０５において、再照合候補を抽出する。本実施の形態では、再照合候補を最大ゆう度リーフノードに対応する語彙とする。他の方法としては、最大ゆう度リーフノードだけでなく、ビーム内に残った累積スコアの上位Ｋ個のリーフノードに対応する語彙をすべて再照合候補とする方法もある。

【００７４】次に、再照合用ツリー展開処理Ｓ０６において、再照合候補の語彙に対して音素マージを行わない第１段～第ｎ段までの音声片ツリーを展開する。この音声片ツリーを再照合用前半音声片ツリー１３とする。再照合用前半音声片ツリー１３は、第１段～第ｎ段で認識語彙が一意に決まる。そこで、第ｎ段の終端ノードにその語彙を割り当てておく。再照合候補が「めぐろ」「めむろ」「ねむろ」の３単語、ｎ＝３であった場合の、再照合用前半音声片ツリー１３の例を図８に示す。

【００７５】本実施の形態では、照合処理Ｓ０２をあらかじめＤＰマッチングを行う際、第１段の始端ノードに対応する入力フレーム位置Ｆｓと、第ｎ段の終端ノードに対応する入力フレーム位置Ｆｅを記憶しておく必要がある。

【００７６】前半再照合処理Ｓ０４では、再照合用前半音声片ツリー１３にしたがって接続した音声片標準パターン１５と、フレームＦｓからフレームＦｅまでの入力音声とを、ＤＰマッチングにより再照合する。再照合の場合は、認識対象語彙が少ないためビームサーチは必ずしも行わなくてもよい。再照合の結果、最も累積スコアの高かった再照合用ツリーの第ｎ段の終端ノードに対応する語彙を認識結果として出力する。

【００７７】なお、ビーム内に残った累積スコアの上位Ｋ個のリーフノードに対応する語彙をすべて再照合候補とする方法の場合には、発声の前半部分のスコア、すなわち再照合の結果求まる入力フレームＦｓからフレームＦｅまでのスコアＳａと、発声の後半部分のスコア、すなわち一回目の照合の結果求まる入力フレームＦｅ＋１から発声の終端フレームまでのスコアＳｂとの和Ｓを、再照合候補の語彙すべてについて求め、Ｓの最も大きい語彙を認識結果とする。

【００７８】本実施の形態では、音素マージを行うのは一律第１段～第ｎ段としたが、すべての段において行っても良い。また、ツリーの密集しているところは深くしたりするなど部分的に変えても良い。再照合を行うのも一律第ｎ段の終端ノードまでではなく、単語が一意に決まるノードまでとしてもよい。すべての段において音素マージを行う場合には一回目の照合では音素マージしていない音声片標準パターン１５を使用する必要はない。

【００７９】また、本実施の形態では、最大ゆう度リーフノードに対応する語彙が１個であった場合は再照合を行わないとしたが、その場合でも、最大ゆう度リーフノードだけでなく、ビーム内に残った累積スコアの上位Ｋ個のリーフノードに対応する語彙をすべて再照合候補としてもよい。

【００８０】以上のように、本実施の形態によれば、第一段から第ｎ段までの音声片について、同じ音素群に属する子音をマージした音素マージ音声片ツリーを用いることにより、一回目の照合における計算量を大幅に削減することができ、再照合を行ったとしても全体の計算量は大幅に削減することができるという効果があります。

【００８１】また、この方法では似た音素を区別しないで認識するため、一回目の照合で正解候補が漏れる可能性が低く、認識性能を劣化させずに計算量を削減することができるという効果があります。

【００８２】さらに、本実施の形態では、再照合は１回目の照合で第１段の始端ノードに対応する入力フレーム位置Ｆｓと第ｎ段の終端ノードに対応する入力フレーム位置Ｆｅを記憶しておき、ＦｓからＦｅまでの間でのみ再照合を行えばよいため、再照合にかかる計算量は非常に少なくてすむという効果があります。

【００８３】（実施の形態２）次に、本発明の実施の形態２の音声認識装置について、図１０のフローチャートを参照しながらその動作を説明する。

【００８４】実施の形態１と異なるのは、再照合用前半音声片ツリー１３が再照合用音声片ツリー１４に、前半再照合処理Ｓ０４が再照合処理Ｓ２１になっていることである。再照合用音声片ツリー１４は、実施の形態１と異なり、第１段～第ｎ段だけではなく、単語終端までを表すツリーになっている。

【００８５】実施の形態２の動作は、ほぼ実施の形態１と同じであるため、異なる部分についてのみ説明する。

【００８６】実施の形態１では、再照合は、１回目の照合のときに音素マージ音声片ツリーの音素マージを行った第１段～第ｎ段に対応していた入力区間についてのみ行ったが、実施の形態２では、発声区間全体について再照合を行う。

【００８７】再照合用ツリー展開処理Ｓ０９では、再照合候補の語彙に対して音素マージを行わない音声片ツリーを展開する。この音声片ツリーを再照合用音声片ツリー１４とする。再照合用音声片ツリー１４は、第１段～第ｎ段までではなく、単語終端までを表す音声片ツリーである。

【００８８】再照合候補が「めぐろ」「めむろ」「ねむろ」の３単語であった場合の、再照合用音声片ツリーの例を図１１に示す。

【００８９】本実施の形態では、照合処理Ｓ０２で、第１段の始端ノードに対応する入力フレーム位置および、第ｎ段の終端ノードに対応する入力フレーム位置を記憶しておく必要はない。

【００９０】再照合処理Ｓ２１では、再照合用音声片ツリー１４にしたがって接続した音声片標準パターン１５と、入力音声の発声開始から発声終了までを、ＤＰマッチングにより再照合する。再照合の場合は、実施の形態１と同様、認識対象語彙が少ないためビームサーチは必ずしも行わなくてよもい。

【００９１】再照合処理Ｓ２１の結果、最も累積スコアの高かった再照合用ツリーのリーフノードに対応する語彙を認識結果として出力する。

【００９２】以上のように、実施の形態２によれば、１回目の照合で第ｎ段の終端ノードに対応する入力フレーム位置が最適な位置ではなかった場合に、発声区間の開始から終了までを再照合することにより、より精密な照合を行うことができるため、実施の形態１に比べさらに認識性能が向上するという効果があります。

【００９３】また、実施の形態２では、第１段の始端ノードに対応する入力フレーム位置Ｆｓと第ｎ段の終端ノードに対応する入力フレーム位置Ｆｅを記憶しておく必要がないため１回目の認識処理およびメモリ容量は実施の形態１に比べ少なくてすむという効果があります。

【００９４】また、実施の形態２のように、発声区間の開始から終了までを再照合する場合には、再照合の距離尺度は１回目のものとまったく違うものを用いてもかまわない。そのため、再照合の際には数単語のみより精密に認識できる方法を用いて、より高い認識性能を得ることもできる。

【００９５】（実施の形態３）次に、本発明の実施の形態３における音声認識装置について、図１２のフローチャートを参照しながらその動作を説明する。

【００９６】実施の形態１と異なるのは、音素マージ音声片ツリー１１がラフ音声片ツリー１２に、音素マージ音声片標準パターン１６がラフ音声片標準パターン１７に、音声片標準パターン１５が精密音声片標準パターン１８になっていること、および判定処理Ｓ０３が不要なことである。

【００９７】精密音声片標準パターン１８は、実施の形態１の音声片標準パターン１５と同じものである。実施の形態３では、ラフ音声片と対比づけるために通常の音声片を精密音声片と呼ぶことにする。

【００９８】ラフ音声片ツリー１２およびラフ音声片標準パターン１７について以下に説明する。ラフ音声片

は、音声片の標準パターンの精度を粗くしたものとして定義する。その方法としては、次の二つが考えられる。

【００９９】一つ目は、ラフ音声片１つの音声片あたりにかかる距離計算量を精密音声片１つあたりにかかる計算量に比べ削減する方法である。具体的には、ラフ音声片標準パターンの、フレーム数を少なくする方法、ガウス分布の混合数を削減する方法、ガウス分布の共分散行列を共通化して共分散行列の種類数を削減する方法などが考えられる。この方法では音声片ツリーの形状は変わらない。

【０１００】二つ目は、認識結果が一意に決まる範囲内で、異なる音韻環境の音声片をマージする方法である。この方法によっても、ツリーのアークとノードが減るため計算量を削減することができる。たとえば、ＶＣは母音部分が同じであれば後続子音が異なっていても１つの音声片にマージするなどが考えられる。この方法では、音声片ツリーの形状が変わることがある。音声片の単位として音素を用いる場合は、音素の前後の音素環境によって異なる音声片とすることが多いが、中心音素が同じ場合には１つの音声片にマージすることにより、ツリーの広がりを大幅に抑えることが可能になる。当然のことながら中心音素が同じであればマージを行っても認識結果は必然的に一意に決まる。

【０１０１】実施の形態３では、標準パターンのフレーム数を削減する方法と、母音部分が同じで後続子音の異なるＶＣをマージする方法の両方を行う。前者は音声片記号の上にバーをつけて表記し、後者は子音部分をアスタリスクで置き換えて表記することとする。

【０１０２】図１３は、図３の基本音声片ツリーを、第１段～第３段（ｎ＝３）までの音声片をラフ音声片としたラフ音声片ツリーである。第４段以降は、基本音声片ツリーと同じである。ツリーの形状は図２と若干変わっている。なお、ＶＣのマージは、後続子音が同じ音素群の場合のみに限っても良い。

【０１０３】ラフ音声片標準パターン１７は、以下のようにあらかじめ学習し求めておく。標準パターンのフレーム数をもとのフレーム数の半分に減らして学習する。さらにＶＣは母音部分が同じ音声片すべての学習データから学習する。たとえば、音声片／ｅ＊／に対する標準パターンは、母音部分が／ｅ／で後続子音が異なる音声片／ｅｍ／，／ｅｎ／，／ｅｇ／，／ｅｂ／，…のすべての学習データから学習することにより得られる。

【０１０４】実施の形態３の動作は、ほぼ実施の形態１と同じであるため、異なる部分についてのみ説明する。

【０１０５】照合処理Ｓ０２では、ラフ音声片ツリー１２にしたがってラフ音声片標準パターン１７および精密音声片標準パターン１８を接続しながら、実施の形態１と同様にして、未知入力音声の特徴パラメータ時系列と標準パターンとの照合を行う。

【０１０６】照合を行った後、再照合候補抽出処理Ｓ０５で、再照合候補を抽出する。本実施の形態では、ビーム内に残った累積スコアの上位Ｋ個のリーフノードに対応するＫ個の語彙を再照合候補とする。実施の形態１と同様に再照合候補に対して再照合用前半音声片ツリー１３を展開し、前半再照合処理Ｓ０４において発声前半部分について精密な音声片標準パターンで照合を行う。

【０１０７】再照合の結果求まる発声の前半部分のスコアＳａと、一回目の照合の結果求まる発声の後半部分のスコアＳｂとの和Ｓを、再照合候補の語彙すべてについて求め、Ｓの最も大きい語彙を認識結果とする。

【０１０８】本実施の形態では、音素マージを行うのは一律第１段～第ｎ段としたが、ツリーの密集しているところは深くしたりするなど部分的に変えても良い。再照合を行うのも一律第ｎ段の終端ノードまででではなく、単語が一意に決まるノードまでとしてもよい。

【０１０９】以上のように、実施の形態３によれば、音声片の標準パターンの精度を粗くしたラフ音声片ツリーを用いることにより、ラフ音声片の照合にかかる計算量が少なくてすむため、一回目の照合における計算量を大幅に削減することができ、再照合を行っても全体の計算量は削減できる。

【０１１０】発声開始直後の計算量の多いところは粗い照合を、発声開始後しばらくしてからの計算量の少ないところは精密な照合をするため効率が良いという効果があります。

【０１１１】（実施の形態４）次に、本発明の実施の形態４の音声認識装置について、図１４のフローチャートを参照しながらその動作を説明する。

【０１１２】実施の形態３と異なるのは、前半再照合処理Ｓ０４が再照合処理Ｓ２１に、再照合用前半音声片ツリー１３が再照合用音声片ツリー１４になっていることである。実施の形態４は、実施の形態３と実施の形態２の組み合わせである。再照合処理Ｓ２１と再照合用音声片ツリー１４は、実施の形態２と同じである。

【０１１３】実施の形態４の動作は、ほぼ実施の形態３と同じであるが、再照合用ツリー展開処理Ｓ０６において再照合用音声片ツリー１４を作成し、音声の前半部分のみ前半再照合処理Ｓ２１において再照合を行うところは、実施の形態２と同じである。

【０１１４】（実施の形態５）次に、本発明の実施の形態５における音声認識装置について、図１５フローチャートを参照しながらその動作を説明する。

【０１１５】実施の形態３と異なるのは、前半再照合処理Ｓ０４、再照合候補抽出処理Ｓ０６、再照合用ツリー展開処理Ｓ０６、再照合用前半音声片ツリー１３が不要なことである。

【０１１６】実施の形態３では、一回目の照合で認識結果は一意に決まるため、再照合を行わなくても認識結果を出力することができる。そこで再照合を行わずに認識結果を出力するとしたのが実施の形態５である。実施の

形態５の動作は、再照合を行わずに一回目の照合の結果をそのまま認識結果とする以外は実施の形態３と同じである。

【０１１７】実施の形態５では、実施の形態３に比べ、再照合を行わないため認識性能は劣化するが計算量は大幅に削減できる。その場合でも音声片の標準パターンの精度を粗くするのは探索空間の広い語頭付近だけであるため、一律に音声片の標準パターンの精度を粗くするよりも効率的な計算量削減が図れる。また、実施の形態５では、再照合の必要がないため、入力音声の特徴パラメータ情報を記憶しておく必要がなく容量も小さくてすむという利点がある。

【０１１８】

【発明の効果】以上のように本発明は、特徴の似ている音素をマージした音声片を用いて照合を行い、認識結果が一意に決まらなかった場合にのみ再照合を行うことにより、認識性能を落とさずに計算量を削減することができる。

【０１１９】また、語頭付近について音声片の標準パターンの精度を粗くしたラフ音声片ツリーを用いて照合を行ったのち、精密な音声片標準パターンを用いて再照合することによって認識性能を落とさずに効率よく計算量を削減することができる。

【０１２０】さらに、音声片の標準パターンの精度を粗くしたラフ音声片ツリーを用いて照合を行い、再照合を行わない場合には、認識性能の劣化を最小限に抑え計算量を大幅に削減することができる。

【０１２１】さらに、１回目の照合で第１段の始端ノードに対応する入力フレーム位置Ｆｓと第ｎ段の終端ノードに対応する入力フレーム位置Ｆｅを記憶しておき、ＦｓからＦｅまでの間でのみ再照合を行う場合には、再照合にかかる計算量を抑えることができる。

【０１２２】また、発声区間の開始から終了までを再照合する場合には、より精密な再照合が行えるため認識性能の劣化が少なくてすみ、１回目の照合方法とまったく違うものを用いてもかまわないため、より精密な手法で再照合を行った場合にはより高い認識性能を得ることもできる。

【図面の簡単な説明】

【図１】本発明の実施の形態におけるコンピュータを用いた音声認識装置の構成図

【図２】本発明の実施の形態１における音声認識装置のフローチャート

【図３】本発明の実施の形態１における基本音声片ツリーを示す図

【図４】本発明の実施の形態１における音声片列を示す図

【図５】本発明の実施の形態１における音素群の定義を説明する図

【図６】本発明の実施の形態１における音声片のマージを説明する図

【図７】本発明の実施の形態１における音素マージ音声片ツリーを示す図

【図８】本発明の実施の形態１における再照合用前半音声片ツリーを示す図

【図９】本発明の実施の形態１における仮想ＤＰ面を説明する図

【図１０】本発明の実施の形態２における音声認識装置のフローチャート

【図１１】本発明の実施の形態２における再照合用音声片ツリーを示す図

【図１２】本発明の実施の形態３における音声認識装置のフローチャート

【図１３】本発明の実施の形態３におけるラフ音声片ツリーを示す図

【図１４】本発明の実施の形態４における音声認識装置のフローチャート

【図１５】本発明の実施の形態５における音声認識装置のフローチャート

【図１６】従来の音声認識装置の構成図

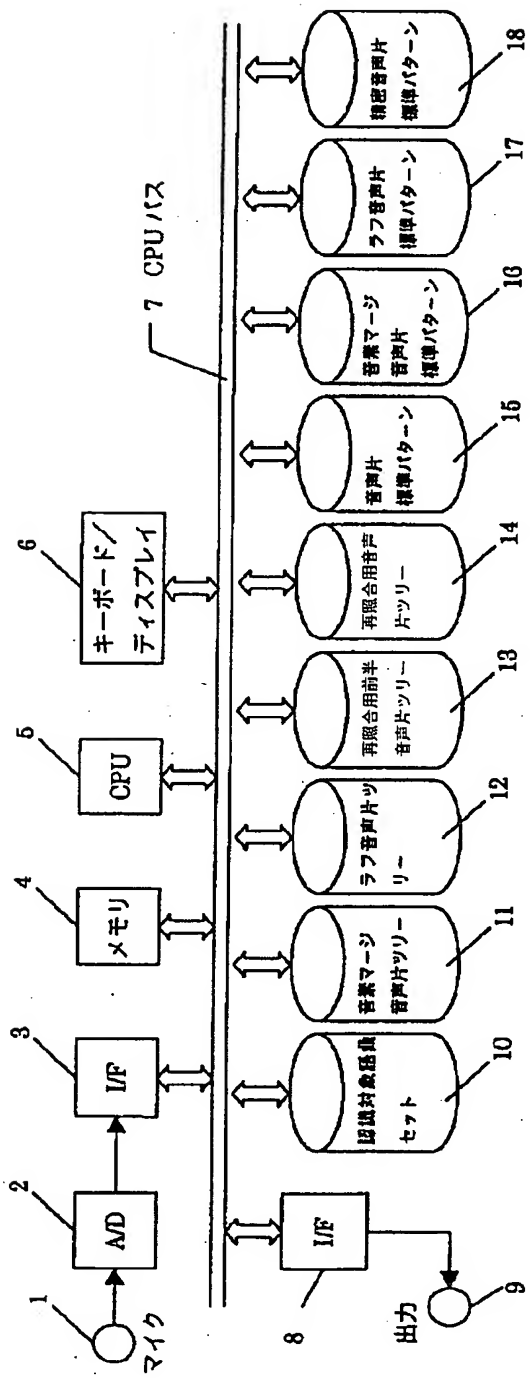
【図１７】従来例における音声認識装置のフローチャート

【図１８】従来例の計算量を説明する図

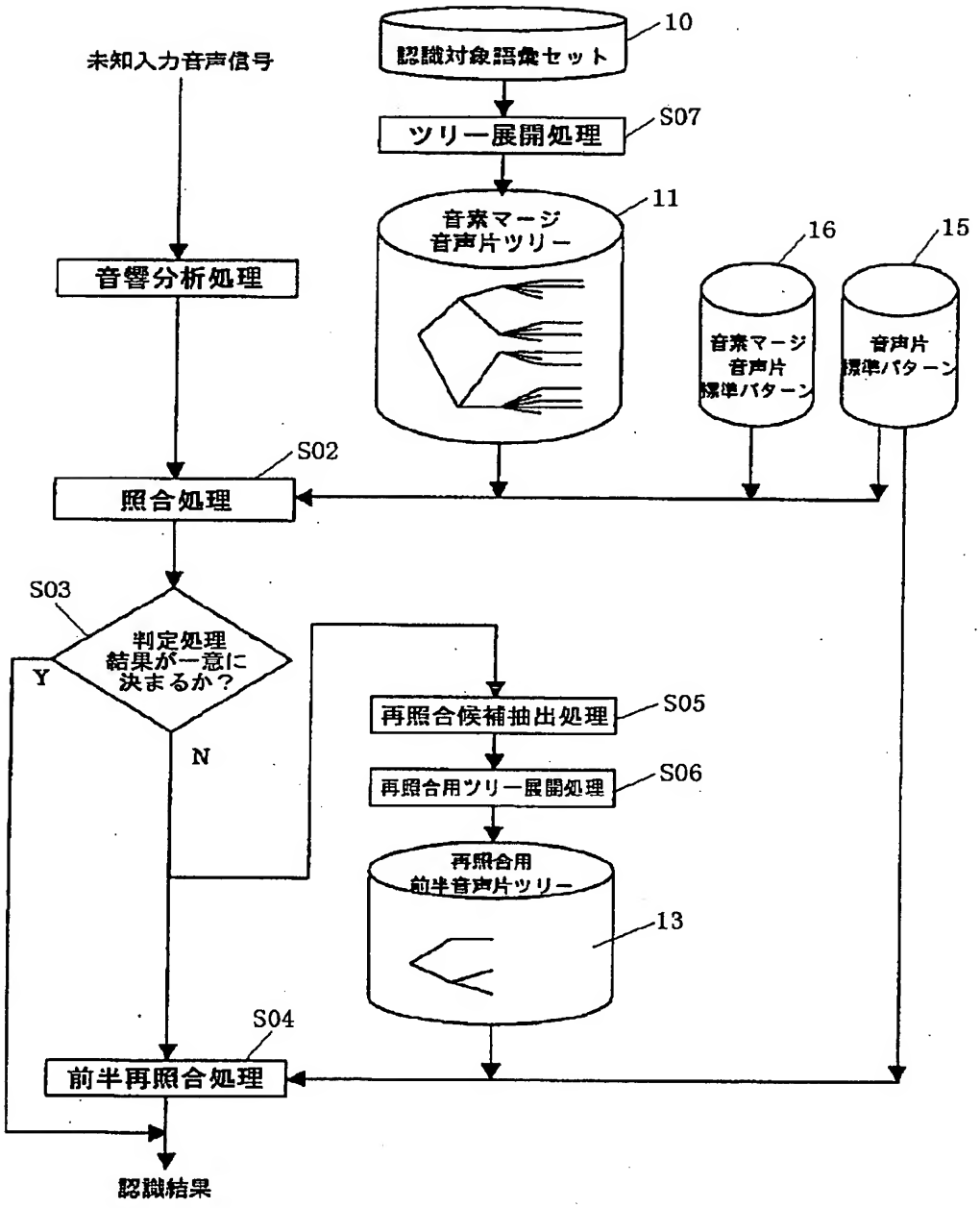
【符号の説明】

１　マイク
２　Ａ／Ｄ
３　インタフェース（Ｉ／Ｆ）
４　メモリ
５　ＣＰＵ
６　キーボード／ディスプレイ
７　ＣＰＵバス
８　Ｉ／Ｆ
９　出力
１０　認識対象語彙セット
１１　音素マージ音声片ツリー
１２　ラフ音声片ツリー
１３　再照合用前半音声片ツリー
１４　再照合用音声片ツリー
１５　音声片標準パターン
１６　音素マージ音声片標準パターン
１７　ラフ音声片標準パターン
１８　精密音声片標準パターン
１９　音声片ツリー

【図１】


1
マイク
2
A/D
3
I/F
4
メモリ
5
CPU
6
キーボード/ディスプレイ
7 CPUバス
8
I/F
9
出力
10
認識対象語彙セット
11
音素マージ音声片ツリー
12
ラフ音声片ツリー
13
再照合用前半音声片ツリー
14
再照合用音声片ツリー
15
音声片標準パターン
16
音素マージ音声片標準パターン
17
ラフ音声片標準パターン
18
精密音声片標準パターン

【図２】

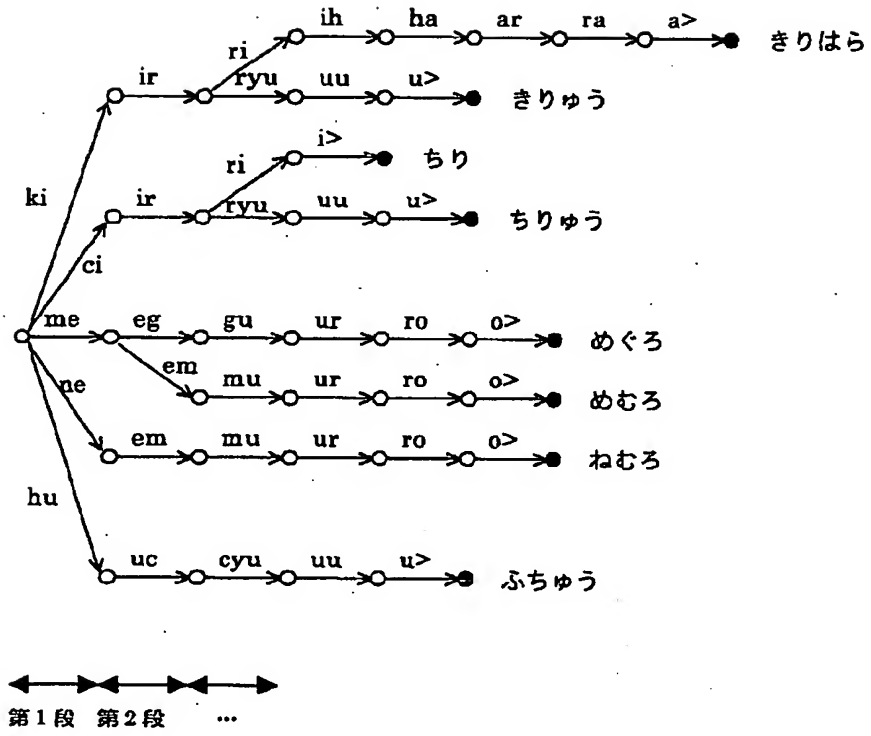


【図５】

無声破裂音・無声破擦音群	{/p/,/t/,/k/,/c/}
無声摩擦音群	{/s/,/h/,/z/}
鼻音群	/m/,/n/}
有性破裂音・流音群	{/b/,/d/,/g/,/r/}

【図３】

基本音声片ツリー



【図４】

音声片列

ki－ir－ri－ih－ha－ra－a>
ki－ir－ryu－uu－u>
ci－ir－ri－i>
ci－ir－ryu－uu－u>
me－eg－gu－ur－ro－o>
me－em－mu－ur－ro－o>
ne－em－mu－ur－ro－o>
hu－uc－cyu－uu－u>

ただし">"は語尾記号

【図６】

CV のマージ例：

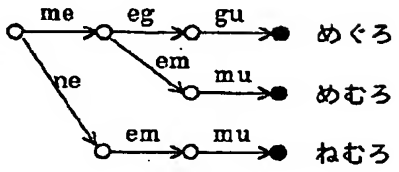
もとの音声片表記		マージ音声片表記
/pi/ , /ti/ , /ki/ , /ci/	→	/{p,t,k,c}i/

VC のマージ例：

もとの音声片表記		マージ音声片表記
/ap/,/at/,/ak/,/ac/	→	/a{p,t,k,c}/

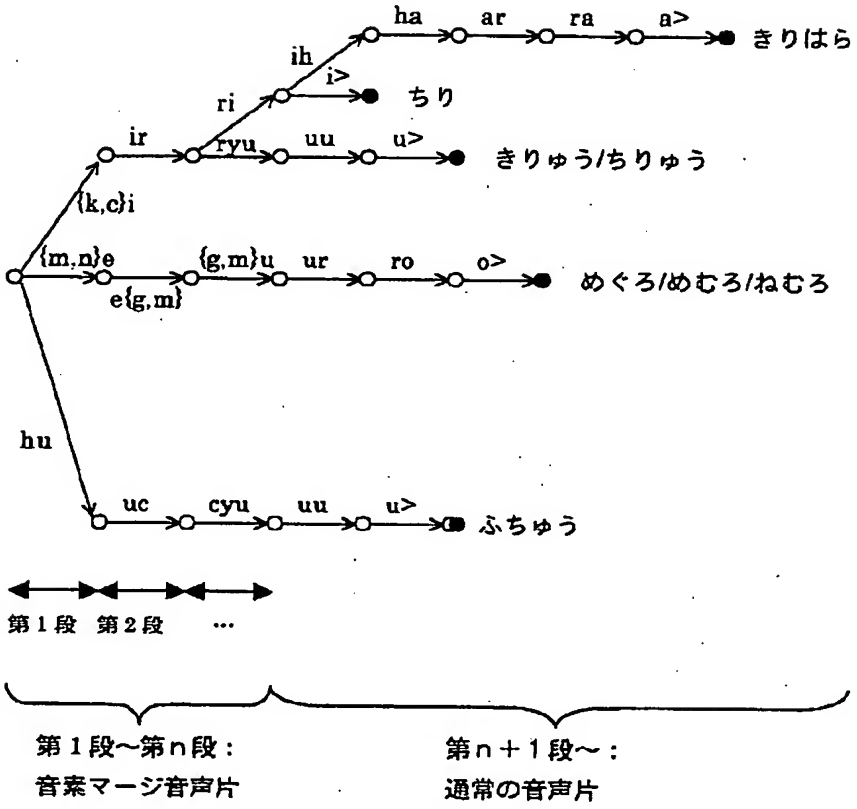
【図８】

再照合用前半音声片ツリー

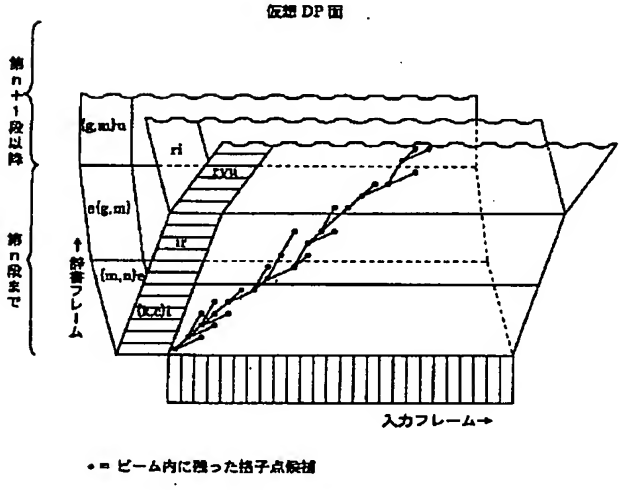


【図７】

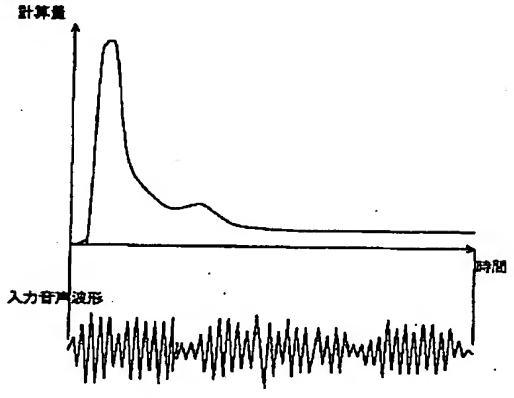
音素マージ音声片ツリー



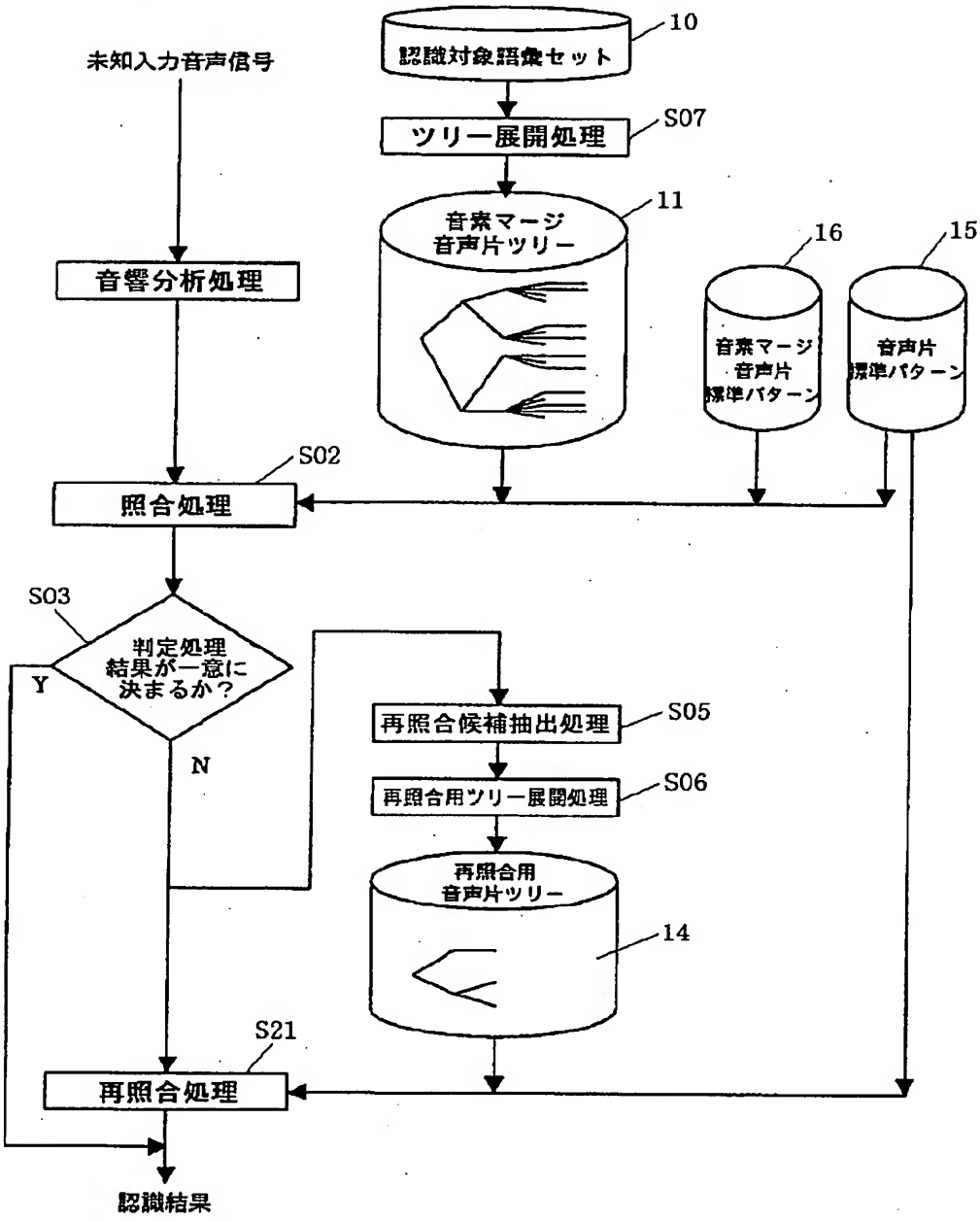
【図９】



【図１８】

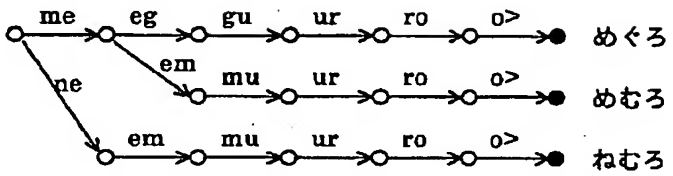


【図１０】

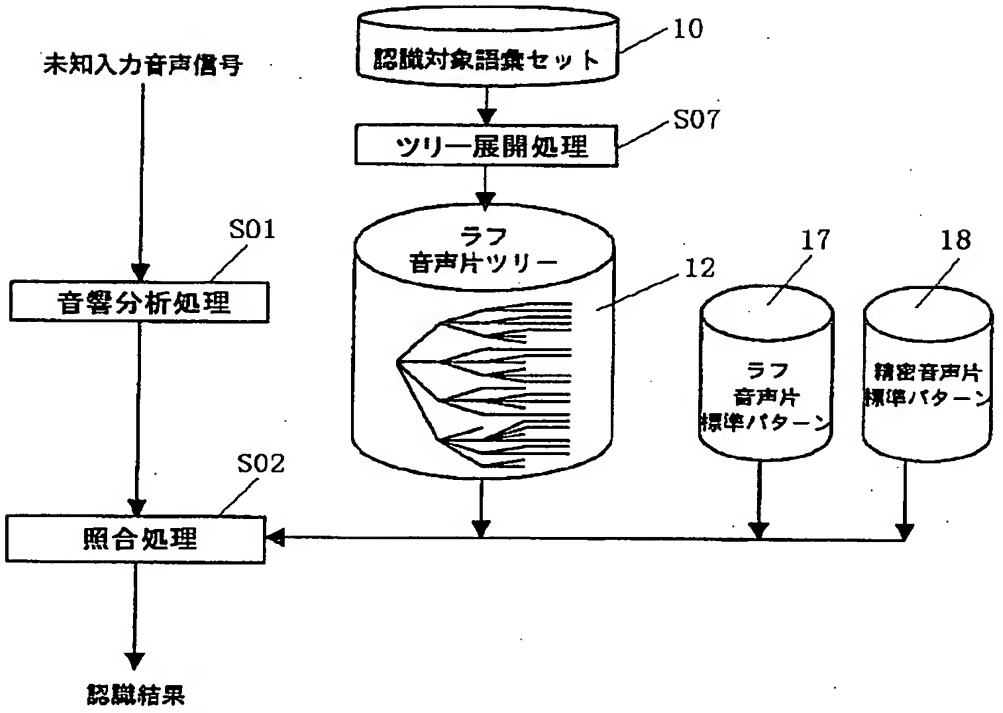

未知入力音声信号
音響分析処理
照合処理
S02
S03
判定処理
結果が一意に
決まるか？
Y
N
認識結果
認識対象語彙セット
10
ツリー展開処理
S07
音素マージ
音声片ツリー
11
音素マージ
音声片
標準パターン
16
音声片
標準パターン
15
再照合候補抽出処理
S05
再照合用ツリー展開処理
S06
再照合用
音声片ツリー
14
S21
再照合処理

【図１１】

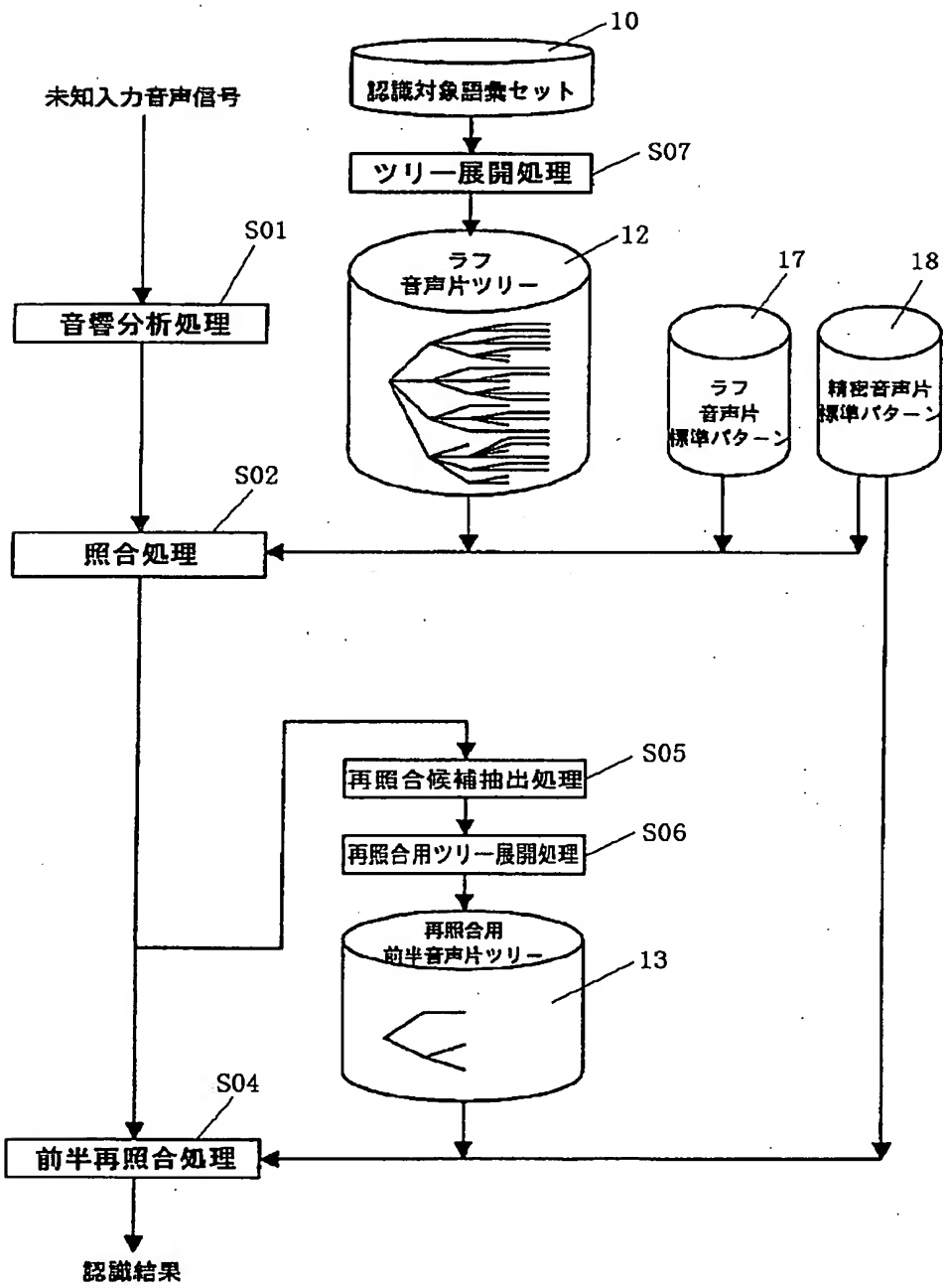
再照合用音声片ツリー



【図１５】

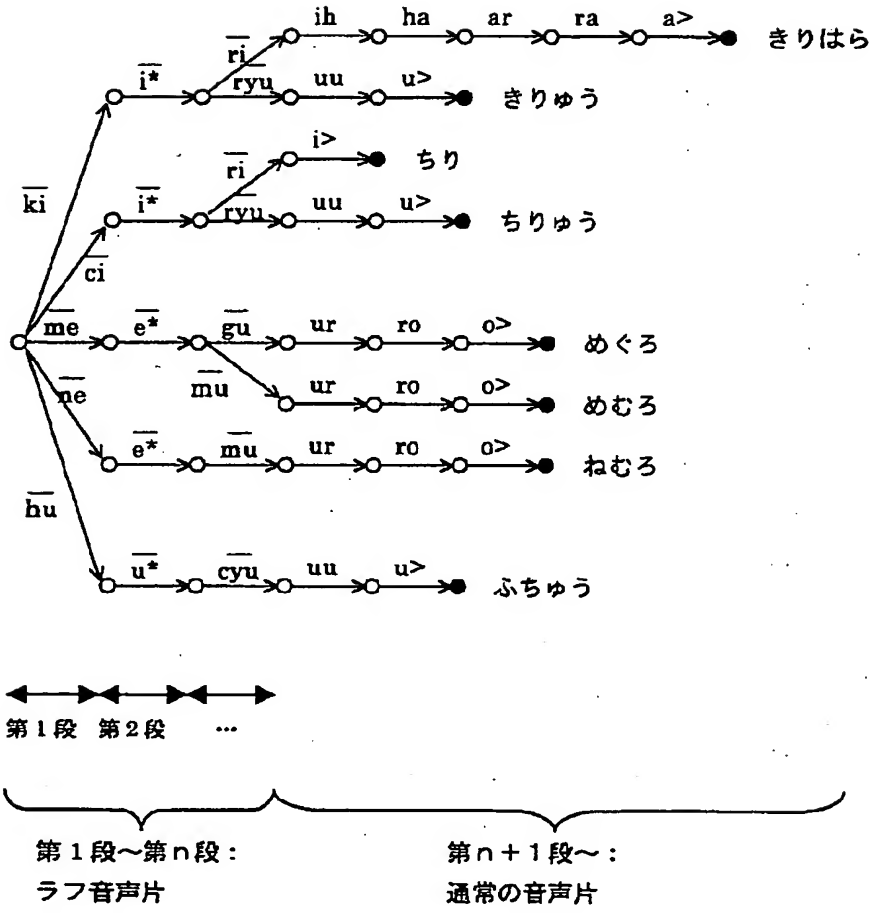


【図１２】


10
認識対象語彙セット
未知入力音声信号
ツリー展開処理
S07
S01
12
ラフ
音声片ツリー
17
18
音響分析処理
ラフ
音声片
標準パターン
精密音声片
標準パターン
S02
照合処理
再照合候補抽出処理
S05
再照合用ツリー展開処理
S06
再照合用
前半音声片ツリー
13
S04
前半再照合処理
認識結果

【図１３】

(a) ラフ音声片ツリー



(b)

ラフ音声片の表記

(1) 標準パターンのフレーム数を削減する場合：

音声片記号の上にバーをつけて表記

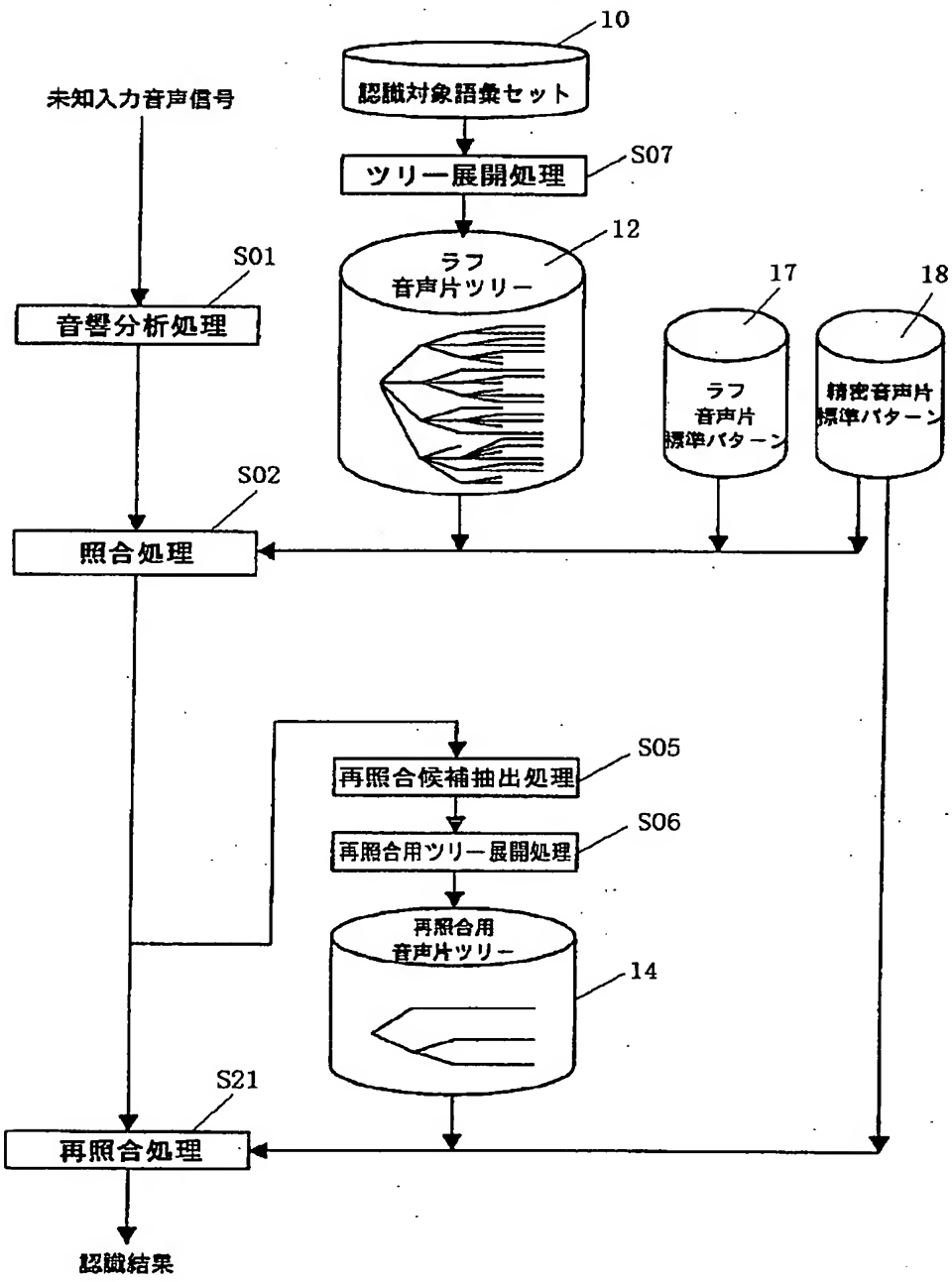
(例) ki → $\overline{\text{ki}}$

(2)母音部分が同じで後続子音の異なる VC をマージする。

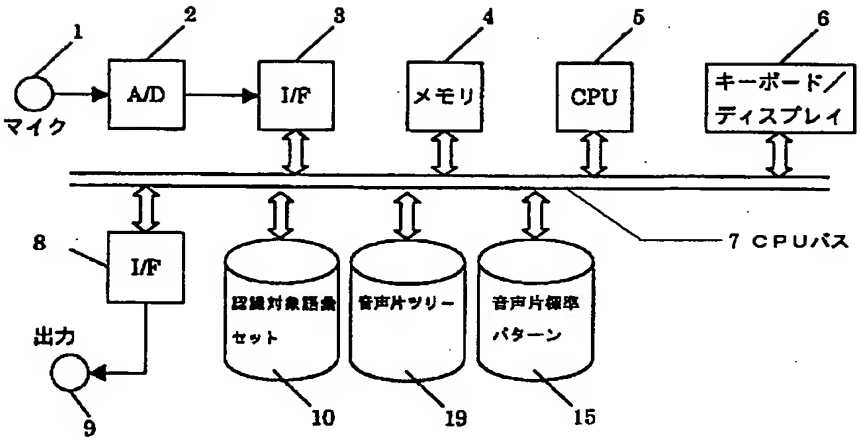
子音部分をアスタリスクに置き換えて表記

(例) em , en , eg , eb , … → e*

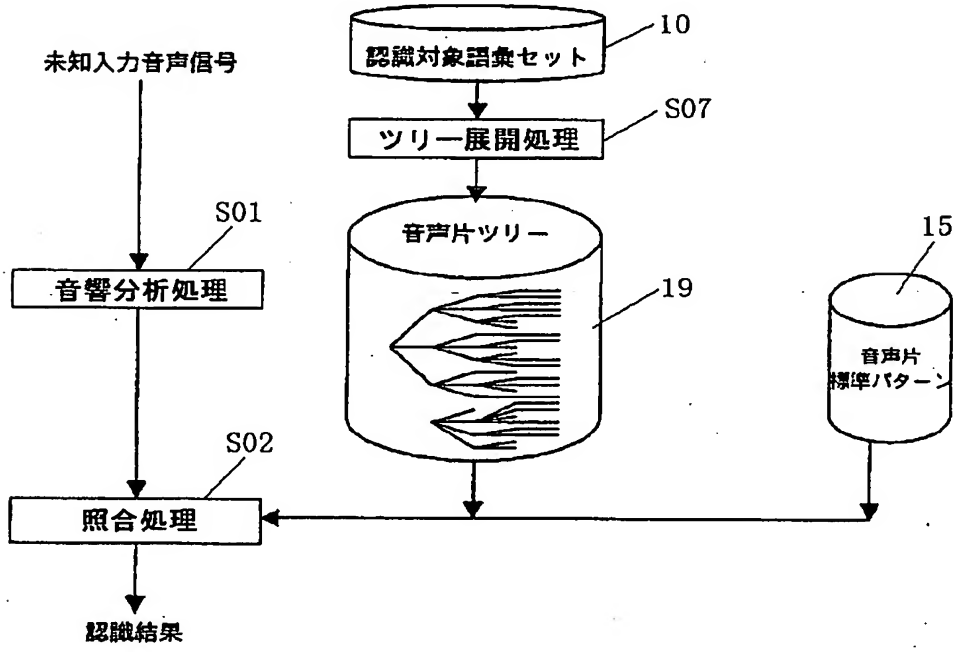
【図１４】


10
認識対象語彙セット
未知入力音声信号
ツリー展開処理
S07
12
ラフ
音声片ツリー
S01
音響分析処理
17
18
ラフ
音声片
標準パターン
精密音声片
標準パターン
S02
照合処理
再照合候補抽出処理
S05
再照合用ツリー展開処理
S06
再照合用
音声片ツリー
14
S21
再照合処理
認識結果

【図１６】


1
2
3
4
5
6
マイク
A/D
I/F
メモリ
CPU
キーボード／
ディスプレイ
7 CPUバス
8
I/F
出力
9
認識対象語彙
セット
音声片ツリー
音声片標準
パターン
10
19
15

【図１７】


未知入力音声信号
認識対象語彙セット
10
S07
ツリー展開処理
S01
音響分析処理
音声片ツリー
19
15
音声片
標準パターン
S02
照合処理
認識結果

[0050] (Embodiment 1)

Fig. 1 is a block diagram of a speech recognition apparatus of Embodiment 1 of the present invention, which will hereinafter be described.

[0051] In Fig. 1, reference numeral 1 denotes a microphone for collecting speech, reference numeral 2 denotes an A/D converter, reference numeral 3 denotes an interface (I/F), reference numeral 4 denotes a memory, reference numeral 5 denotes a CPU, reference numeral 6 denotes a keyboard/display, reference numeral 7 denotes a CPU bus, reference numeral 8 denotes an I/F, reference numeral 9 denotes an output, reference numeral 10 denotes a recognition target dictionary set, reference numeral 11 denotes a phoneme merged speech segment tree, reference numeral 12 denotes a rough speech segment tree, reference numeral 13 denotes a first half speech segment re-collation tree, reference numeral 14 denotes a speech segment re-collation tree, reference numeral 15 denotes a speech segment standard pattern, reference numeral 16 denotes a phoneme merged speech segment standard pattern, reference numeral 17 denotes a rough speech segment standard pattern, and reference numeral 18 denotes an accurate speech segment standard pattern.

[0052] First, the phoneme merged speech segment tree 11

corresponding to a recognition dictionary in Embodiment 1 will be described with reference to Figs. 3, 4 and 5.

[0053] As a unit of a standard pattern, a phoneme segment, a phoneme, a syllable, CV/VC (a consonant + a vowel/ a vowel + a consonant), VCV, CVC and the like are conceivable. These minimum recognition units are referred to as speech segments. In the present embodiment, CV representing from the start of a consonant to the center of a vowel, VC representing from the center of a vowel to the end of the vowel, and VV representing from the center of a vowel to the center of a vowel serve as base units. Although VC includes only a vowel section, it is defined differently depending on the subsequent consonant.

[0054] For example, assume that the recognition target vocabulary includes 8 words: "kirihara", "kiryu", "chiri", "chiryu", "meguro", "memuro", "nemuro" and "fuchu", and when they are represented by speech segment sequences, the results are as shown in Fig. 4.

[0055] Fig. 3 is a diagram showing the above results in a simple tree structure. In the present embodiment, this is defined as a base speech segment tree. This is the same as the speech segment tree used in the prior art. Although speech segments are allocated to arcs here, they may also be allocated to nodes. Each node corresponding to an end of a vocabulary is made recognizable as the end of the

vocabulary. Such a node is defined as a leaf node. In Fig. 3, the leaf nodes are represented by solid circles. Further, the depth of the tree is to be counted from its root (e.g., a first step, a second step ···).

[0056] In the phoneme merged speech segment tree 11, phonemes from a first to nth steps of the base phoneme tree are merged, whereby expansion of initial phonemes of words in the tree is suppressed. The phoneme merged speech segment tree is the same as the base speech segment tree subsequent from a (n+1)th step.

[0057] Phoneme merging from the first to nth steps is performed by the following method. There are only five vowels in Japanese so that it is relatively easy to distinguish them. On the other hand, it is very difficult to distinguish consonants because the number of categories is large. Consonants are collectively merged for each phoneme group (e.g., unvoiced stop consonant, fricative consonant, voiced stop consonant groups and the like), and consonants within the same phoneme group are not distinguished. That is, in each consonant, phonemes are not distinguished, and they are distinguished only by the phoneme group such as the unvoiced stop consonant or fricative consonant group. This means that "kiryu" and "chiryu", which are different by one letter in initial phoneme, are to be collated without any distinction.

[0058] Since merging is performed between consonants having similar acoustic features in the same phoneme group, there is little error due to merging, and moreover, distinction between different phoneme groups is easy because their acoustic features are very different. Therefore, a correct candidate is hardly pruned so that deterioration of recognition performance is little. In the present embodiment, the consonants are divided into four categories as shown in Fig. 5.

[0059] By merging phonemes, speech segments are also merged. Merging occurs to CV when its subsequent vowel is the same, and merging occurs to VC when its antecedent phoneme is the same. A speech segment obtained by merging phonemes for each phoneme group is defined as a phoneme merged speech segment. Examples of the merging method of the phoneme merged speech segment and its notation are shown in Fig. 6.

[0060] Of the base speech segment tree, speech segments from the first to nth steps are converted to phoneme merged speech segments, whereby arcs to which the same phoneme merged speech segment is allocated are merged making it possible to obtain a tree with a reduced expansion in the vicinity of initial phonemes of words. This is the phoneme merged speech segment tree.

[0061] Fig. 7 is a phoneme merged speech segment tree

obtained by merging speech segments from the first to third steps (n=3) of the base speech segment tree of Fig. 3 and converting them to phoneme merged speech segments. In the phoneme merged speech segment tree of Fig. 7, expansion of initial phonemes of words in the tree is suppressed, compared with the speech segment tree of Fig. 3. Supposing that n is 1, first speech segments of the initial phonemes of the words are merged, and supposing that n is ∞, all the speech segments are to be merged. It is efficient if the size of n is determined to an extent that the quantity of calculation is within the real-time processing. In the phoneme merged speech segment tree, a plurality of words may be allocated to a leaf node as in the case of "kiryu" and "chiryu".

[0062] Next, operation of a speech recognition apparatus in Embodiment 1 of the present invention will be described with reference to a flowchart of Fig. 2.

[0063] In Fig. 2, the speech segment standard pattern 15 is found for each speech segment by learning in advance learned data in which a lot of speakers vocalized. The phoneme merged speech segment standard pattern 16 is found by learning all the learned data of speech segments that are merged. For example, a standard pattern for the speech segment/ {p, t, k, c} i/ can be obtained by learning all the learned data of /pi/, /ti/, /ki/, and /ci/. This is to

be found in advance for each of all the phoneme merged speech segments.

[0064] In the present embodiment, assuming that the appearance probability of a feature parameter vector can be approximated by a sum of a plurality of Gaussian distributions, (which is referred to as a mixed distribution), average value vectors and covariance matrices in the Gaussian distribution are found from the learned data for each frame of the standard pattern.

[0065] The phoneme merged speech segment tree 11 is made in advance from the recognition target vocabulary set 10 in development processing of the tree S07.

[0066] First, in acoustic analysis processing S01, an inputted unknown speech signal is converted to D feature parameters for each analysis time (hereinafter referred to as frame). Examples of the feature parameters include LPC cepstrum coefficient and LPC mel-cepstrum coefficient in accordance with linear prediction analysis; mel-LPC cepstrum coefficient in accordance with mel-linear prediction analysis; mel-frequency cepstrum coefficient (MFCC) in accordance with mel-scale filter bank; and the like, and any feature parameter may be used as long as it is suited for speech recognition.

[0067] In collation processing S02, in accordance with the phoneme merged speech segment tree 11, a feature

parameter time series for the unknown inputted speech from the acoustic analysis processing S01 is collated with that for the standard pattern while connecting the phoneme merged speech segment standard pattern 16 and the speech segment standard pattern 15. The collation is performed by DP matching using input frame synchronous beam search. Since the method for DP matching and the method for beam search are the same as those of the conventional example, their description is omitted. The collation is referred to as first collation.

[0068] Although the phoneme merged speech segment tree 11 is to be made in advance in the present embodiment, it may also be dynamically developed while performing beam search DP.

[0069] A conceptual diagram as to DP using input frame synchronous beam search is shown in Fig. 9. In Fig. 9, an axis of abscissas represents a frame of inputted speech, while an axis of ordinates represents a frame of a speech segment standard pattern connected in accordance with the phoneme merged speech segment tree. The axis of ordinates serving as a dictionary is in a tree shape. DP matching between the inputted speech and the tree-shaped dictionary is to calculate scores while finding an optimal path of the inputted speech pattern and the standard pattern on the tree-shaped DP screen. In the tree-shaped DP screen,

phonemes from the first to nth steps have been merged, so that the expansion of branches is suppressed.

[0070] In DP matching, pruning DP paths is carried out synchronously with the input frame by beam search. Since the number of candidate lattice points remaining in a beam is far smaller than the number of all the lattice points on the DP screen, the DP screen is not necessary for an actual memory, which is a virtual one.

[0071] After a while from the start of vocalization, cumulative scores of the DP paths in the dictionary, which are not similar to the speech content, become small values enough to be pruned and thus the number of candidate lattice points is drastically reduced. Therefore, the reduction of the number of candidate lattice points so far leads to the reduction of the total calculation quantity. As in the first embodiment, by suppressing the expansion in the vicinity of initial phonemes of words of the tree, the number of candidate lattice points remaining in the beam in the vicinity of the start of speech can greatly be reduced.

[0072] In determination processing S03, a leaf node (maximum likelihood leaf node) is found, and determination of whether or not a word corresponding thereto is uniquely determined is performed. If it is uniquely determined (Y), namely in the case where there is only one word corresponding to the maximum likelihood leaf node, the word

is outputted as a recognition result. If it is not uniquely determined (N), namely in the case where there are a plurality of words corresponding to the maximum likelihood leaf node, the recognition result is determined in the following method.

[0073] In re-collation candidate extraction processing S05, re-collation candidates are selected. In the present embodiment, the re-collation candidates are those vocabularies corresponding to the maximum likelihood leaf node. As another method, there is also a method in which all the vocabularies corresponding not only to the maximum likelihood leaf node but also to the top K leaf nodes of the cumulative scores remaining in the beam.

[0074] Next, in re-collation tree development processing S06, a first to nth steps of a speech segment tree, where phoneme merging is not performed with respect to re-collation candidate vocabularies, are developed. This speech segment tree serves as a first half speech segment re-collation tree 13. In the first half speech segment re-collation tree 13, a recognition vocabulary is uniquely determined in the first to nth steps. Thus, the vocabulary is allocated to an end node of the nth step. An example of the first half speech segment re-collation tree 13, where re-collation candidates are three words of "meguro", "memuro" and "nemuro", i.e., n=3, is shown in Fig. 8.

[0075] In the present embodiment, in performing the collation processing S02 by DP matching beforehand, it is required to store an input frame position Fs corresponding to a start node of the first step and an input frame position Fe corresponding to an end node of the nth step.

[0076] In first half re-collation processing S04, an input speech from the frame Fs to the frame Fe is re-collated with a speech segment standard pattern 15 connected in accordance with the first half speech segment re-collation tree 13 by DP matching. In the case of re-collation, since the recognition target vocabulary is small, beam search is not necessarily required to be performed. As a result of the re-collation, a word having the highest cumulative score, which corresponds to the end node of the nth step of the re-collation tree, is outputted as a recognition result.

[0077] In the case of the method in which all the words corresponding to the top K leaf nodes of the cumulative scores remaining in the beam serve as re-collation candidates, a sum S of a score of a first half of vocalization, namely a score Sa from an input frame Fs to a frame Fe, which is found as a result of the re-collation, and a score of a second half of vocalization, namely a score Sb from an input frame Fe+1, which is found as a result of the first collation, to an end frame of

vocalization is found for each re-collation candidate vocabulary. Then, a word with the largest S is determined as a recognition result.

[0078] In the present embodiment, although phoneme merging is performed collectively from the first to nth steps, it may be performed in all the steps. Further, a portion of the tree, which is densely divided, may be partially modified by further dividing and the like. Instead of collectively performing re-collation to the end node of the nth step, re-collation may be performed to the node in which the word is uniquely determined. In the case where phoneme merging is performed in all the steps, it is not necessary to use the speech segment standard pattern 15 in which phonemes are not merged.

[0079] In the present embodiment, if there is only one words corresponding to the maximum likelihood leaf node, re-collation is not performed. In this case also, not only the maximum likelihood leaf node but also all the words corresponding to the top K leaf nodes of the cumulative scores may serve as re-collation candidates.

[0080] As described above, according to the present embodiment, by using the phoneme merged speech segment tree in which consonants belonging to the same phoneme group are merged as to speech segments from the first to nth steps, the quantity of calculation in the first collation can

greatly be reduced, and, even if re-collation is performed, the total quantity of calculation can greatly be reduced.

[0081] Further, in this method, since similar phonemes are recognized without making any distinction, the possibility that correct candidates are failed to be selected is low. Thus, the quantity of calculation can be reduced without deteriorating recognition performance.

[0082] Furthermore, in the present embodiment, storing the input frame position Fs corresponding to the start node of the first step and the input frame position Fe corresponding to the end node of the nth step, re-collation is performed only between the Fs and the Fe, so that the quantity of calculation required for re-collation is so small.

[0083] (Embodiment 2)

Next, operation of a speech recognition apparatus according to Embodiment 2 of the present invention will be described with reference to a flowchart of Fig. 10.

[0084] What differs from Embodiment 1 is that the first half speech segment re-collation tree 13 and the first half re-collation processing S04 are changed to a speech segment re-collation tree 14 and re-collation processing S21, respectively. Different from Embodiment 1, the re-collation speech segment tree 14 is a tree representing not only from a first to nth steps but also to an end step of a

word.

[0085] Since operation of Embodiment 2 is almost the same as that of Embodiment 1, only different parts will be described.

[0086] In Embodiment 1, re-collation is performed only on the input section corresponding to the first to nth steps, where phoneme merging of the phoneme merged speech segment tree was performed in the first collation. On the other hand, in Embodiment 2, re-collation is performed on all the vocalization section.

[0087] In re-collation tree development processing S09, a speech segment tree, where phoneme merging is not performed on re-collation candidate words, is developed. This speech segment tree is referred to as the re-collation speech segment tree 14. The re-collation speech segment tree 14 is a speech segment tree representing not from the first to nth steps but to the end step of the word.

[0088] An example of the re-collation speech segment tree, where re-collation candidates are three words, i.e., "meguro", "memuro" and "nemuro", will be described in Fig. 11.

[0089] In the present embodiment, in collation processing S02, it is not required to store an input frame position corresponding to a start node of the first step and an input frame position corresponding to an end node of

the nth step.

[0090] In re-collation processing S21, an input speech from the start of vocalization to the end of vocalization is re-collated with a speech segment standard pattern 15 connected in accordance with the re-collation speech segment tree 14 by DP matching. In the case of re-collation, since the recognition target vocabulary is small as in Embodiment 1, it is not necessarily required to perform beam search.

[0091] As a result of the re-collation processing S21, a vocabulary corresponding to a leaf node of the re-collation tree, which had the highest cumulative score, is outputted as a recognition result.

[0092] As described above, according to Embodiment 2, in the case where the input frame position corresponding to the end node of the nth step was not an optimal position in the first collation, by performing re-collation from the start to the end of the vocalization section, more accurate collation can be performed. Thus, recognition performance is further improved compared with Embodiment 1.

[0093] In Embodiment 2, since it is not required to store the input frame position Fs corresponding to the start node of the first step and the input frame position Fe corresponding to the end node of the nth step, the first recognition processing and memory capacity are less than

those of Embodiment 1.

[0094] As in Embodiment 2, in the case where the vocalization section is re-collated from its start to end, a distance scale completely different from the one used for the first collation may be used. Therefore, in performing re-collation, higher recognition performance can also be achieved using a method by which higher recognition can be achieved with only several words.

[0095] (Embodiment 3)

Next, operation of a speech recognition apparatus according to Embodiment 3 of the present invention will be described with reference to a flowchart of Fig. 12

[0096] What differs from Embodiment 1 is that the phoneme merged speech segment tree 11, the phoneme merged speech segment standard pattern 16, the speech segment standard pattern 15 are changed to a rough speech segment tree 12, a rough speech segment standard pattern 17, an accurate speech segment standard pattern 18, respectively, and that the determination processing S03 is not required.

[0097] The accurate speech segment standard pattern 18 is the same as the speech segment standard pattern 15 of Embodiment 1. In Embodiment 3, a normal speech segment is referred to as an accurate speech segment in order to contrast it with a rough speech segment.

[0098] The rough speech segment tree 12 and the rough

speech segment standard pattern 17 will hereinafter be described. The rough speech segment is defined as the one that is obtained by reducing the accuracy of the speech segment standard pattern. The following two methods are conceived as a method therefor.

[0099] The first is a method in which the quantity of distance calculation required per one rough speech segment is reduced, compared with that required per one accurate speech segment. Specifically, a method for reducing the number of frames of the rough speech segment pattern, a method for reducing the mixture number of the Gaussian distribution, a method for making the covariance matrices of the Gaussian distribution common to reduce the number of kinds of covariance matrices, and the like are conceived. In this method, the shape of the speech segment tree is not changed.

[0100] The second is a method for merging speech segments of different phonological environment in a range in which the recognition result is uniquely determined. By this method also, since arcs and nodes of the tree are reduced, the quantity of calculation can be reduced. For example, the case where VCs, whose vowel portions are the same, are merged with one speech segment even if their subsequent consonants are different and the like are conceived. In this method, the shape of the speech segment

tree may be changed. In the case where phonemes are used as units of the speech segment, speech segments are determined to be different in many cases depending on phonemic environment around them, but in the case where they have the same center phoneme, they are merged with one speech segment, whereby the expansion of the tree can greatly be suppressed. As a matter of course, if they have the same center phoneme, the recognition result is naturally and uniquely determined even if merging is performed.

[0101] In Embodiment 3, both of the method for reducing the number of frames of the standard pattern and the method in which VCs having the same vowel portion and different subsequent consonants are merged are used. The former is represented by affixing a bar on a speech segment symbol, while the latter is represented by replacing a consonant portion with an asterisk.

[0102] Fig. 13 is a rough speech segment tree obtained by converting speech segments from the first to third steps (n=3) of the base speech segment tree of Fig. 3 to rough speech segments. Speech segments after the fourth step are the same as those of the base speech segment tree. The shape of the rough speech segment tree is slightly different from Fig. 2. VC-merging may also be limited to the only case where the subsequent consonant belongs to the

same phoneme group.

[0103] The rough speech segment standard pattern 17 is provided in advance by learning. Based on the number of frames of the standard pattern, learning is performed by reducing the number of frames of the standard pattern to a half of the number of original frames. Moreover, as to VCs, learning is performed from all the learned data of the speech segments having the same vowel portion. For example, the standard pattern for the speech segment /e*/ can be obtained by learning all the learned data of the speech segments whose vowel portions are /e/ and the subsequent consonants are different, i.e., /em/, /en/, /eg/, /eb/, ···.

[0104] Operation of Embodiment 3 is almost the same as that of Embodiment 1 and thus only different portions will be described.

[0105] In collating processing S02, in the same manner as in Embodiment 1, a feature parameter time series of an unknown inputted speech is collated with that of the standard pattern while connecting the rough speech segment standard pattern 17 and the accurate speech segment standard pattern 18 in accordance with the rough speech segment tree 12.

[0106] After performing the collation, re-collation candidates are extracted in re-collation candidate extraction processing S05. In the present embodiment, K

vocabularies corresponding to the top K leaf nodes of the cumulative scores serve as re-collation candidates. In the same manner as in Embodiment 1, the first half speech segment re-collation tree 13 is developed for each re-collation candidate, and a first half of vocalization is collated with the accurate speech segment standard pattern in the first half re-collation processing S04.

[0107] A sum S of a score Sa of a first half of vocalization, which is found as a result of the re-collation, and a score Sb of a second half of vocalization, which is found as a result of the first re-collation, is found for all the candidate vocabularies. Then, a vocabulary with the largest S is determined as a recognition result.

[0108] In the present embodiment, although phoneme merging is performed collectively from the first to nth steps, a portion of a tree, which is densely divided, may be partially modified by further dividing and the like. Instead of collectively performing re-collation to the end node of the nth step, re-collation may be performed until the node in which the word is uniquely determined.

[0109] As described above, according to Embodiment 3, by using the rough speech segment tree whose accuracy of the speech segment standard pattern is reduced, the calculation quantity required for the re-collation of the rough speech

segments is small and thus the calculation quantity in the first collation can greatly be reduced. Even if re-collation is performed, the total calculation quantity can be reduced.

[0110] Since rough collation is performed in a part where the calculation quantity is large, while accurate collation is performed in a part where the calculation quantity is small a little while after the start of vocalization, the efficiency is high.